

Mikko Kangassalo

# **THE EPISTEMIC CONDITION FOR MORAL RESPONSIBILITY**

An Examination of the Searchlight View,  
George Sher's Alternative,  
and a Pragmatic View

Faculty of Social Sciences  
Master's thesis  
May 2019

# ABSTRACT

Kangassalo, Mikko : The Epistemic Condition for Moral Responsibility – An Examination of the Searchlight View, George Sher's Alternative, and a Pragmatic View

Master's thesis

Tampere University

Master's programme in philosophy

May 2019

---

In recent years, there has been a lot of *moral outrage* on social media. Consequently, many seemingly unfruitful as well as misguided assignments of negative moral responsibility have been made by various moral agents and groups of agents. It seems that in public discourse moral responsibility is taken for granted in some implicit form, and the related outrage is dispensed in a very un-meditated fashion. This is well represented in the increased political polarization in the United States and many countries in Europe during the last few years and decades. It is also present in communication about vastly important scientific topics, such as climate change. These phenomena suggest that we ought to think more carefully about the role of knowledge in assigning responsibility, and our relevant ways of thinking and acting.

The knowledge requirement or the *epistemic condition* is, along with the freedom or control requirement, one of the two classic necessary conditions for moral responsibility, first described by Aristotle in the *Nicomachean Ethics*. Lately, a prolific novel discussion about the condition has emerged. Motivated by the desire to examine where our interpersonal communication may be going wrong, the research question driving this thesis is: how should we think about the knowledge requirement of moral responsibility?

Some recent discussion is critically examined. This is done primarily via George Sher's examination of what he calls the *searchlight view*, his expansion of it via his *full account of responsibility's epistemic condition* (FEC), and some answers that his account has generated. Sher argues that unlike the searchlight view suggests, there are situations where a moral agent can be responsible without awareness of morally relevant aspects of their situation. In answers to Sher, special attention is given to critical reviews by Michael J. Zimmerman and Angela Smith, whom Sher implicates as some of his contemporary opponents. The contemporary discussion is also more broadly outlined, with five main positions introduced.

While Sher's argument is intriguing and, in many parts, likely to appeal to many intuitively, considerable reservations are presented. In the current state of the discussion, there is a lot of disagreement on what kind of emphasis knowledge should be understood to have in moral responsibility. The example cases where Sher would have us intuitively see responsibility without awareness are found to be not so straightforward.

Rather than merely trying to entangle our possibly incompatible and conflicting intuitions, an empirically informed pragmatic approach to the question of the role of knowledge in moral responsibility is suggested. An original pragmatic argument is presented for the importance of paying attention to the target agent's awareness and level of knowledge when evaluating and communicating with them. The argument suggests that at least in some situations or topics it stands to reason that not respecting the knowledge requirement in a way that would seem incompatible with Sher's account would lead to a moral paradox as well as to unpragmatic results in public communication. The resulting normative, consequentialist view – *the pragmatic view* – is noted to go against the prevailing descriptive, merit-based discussion about the epistemic condition, and its home is suggested to be located in *fusion* or *cosmopolitan virtue ethics*.

In addition to a theoretical description of the pragmatic view, also a concrete applied answer to the research question and an approximately corresponding illustrative flowchart are formulated. According to the current iteration of the pragmatic view: moral outrage online and offline could be mitigated, public and private communication enhanced, flow of (morally relevant) knowledge in society facilitated, more constructive media conventions fostered, and political polarization reduced via cultivating the presented virtuous, pragmatic heuristic in our habits of agent evaluation and assigning moral responsibility. Further development of the pragmatic view is encouraged, and its application in addition to our personal lives is especially recommended for the contexts of artificial intelligence and social media platform design and design ethics.

Keywords: moral responsibility, epistemic condition, moral outrage, social media, political polarization, science communication, metacognition, moral psychology, virtue ethics, pragmatism

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

# TIIVISTELMÄ

Kangassalo, Mikko : Moraalisen vastuun episteeminen ehto – Tarkastelussa valonheitinnäkemyks, George Sherin vaihtoehto sekä pragmaattinen näkökulma  
Maisterintutkielma  
Tampereen yliopisto  
Filosofian maisteriopinnot  
Toukokuu 2019

---

Viime vuosina sosiaalisessa mediassa on esiintynyt paljon *moraalista tyrmistystä*. Sen seurauksena erinäiset moraaliset agentit ja ryhmät ovat asettaneet toisiaan näennäisen hedelmättömästi, kuin myös harhaanjohtetusti, negatiiviseen moraaliseen vastuuseen erinäisistä asioista. Julkisessa diskurssissa moraalinen vastuu vaikutetaan otettavan annettuna jossain implisiittisessä muodossa ja tähän liittyvää tyrmistystä jaetaan hyvin harkitsemattomin tavoin. Tämä on hyvin edustettuna viime vuosien ja vuosikymmenten aikana tapahtuneessa poliittisen polarisaation nousussa Yhdysvalloissa ja monissa Euroopan valtioissa. Sama ongelma on läsnä myös kommunikoitaessa hyvin tärkeistä ajankohtaisista tieteellisistä aiheista, kuten ilmastonmuutoksesta. Nämä ilmiöt vihjaavat, että meidän tulisi huolellisemmin ajatella tiedon roolia vastuuseen asettamisessa sekä relevantteja tapojamme ajatella ja toimia.

Tietoehto tai *episteeminen ehto* on vapaus- tai kontrolliehdon ohella toinen kahdesta klassisesta moraalisen vastuun välttämättömästä ehdosta, jotka Aristoteles alkujaan esitteli teoksessaan *Nikomakhoksen etiikka*. Viime aikoina kyseinen ehto on synnyttänyt satoja, uutta keskustelua. Motivoituneena halusta selvittää nykyisiä ihmistenväliseen kommunikaatioon liittyviä kompastuskiviämme, tutkielmaa ohjaava tutkimuskysymys on: miten meidän tulisi ajatella moraalisen vastuun tietoehtoa?

Tutkin kriittisesti osaa viimeaikaisesta keskustelusta. Tämä tapahtuu ensisijaisesti George Sherin esittämien tarkastelujen ja hänen saamiensa vastausten valossa. Sher esittää kriittisen selonteon *valonheitinnäkemyksestä*, laajentaen sitä hänen *episteemisen ehdon täyden selonteonsa* kautta. Hän esittää, että toisin kuin valonheitinnäkemyks antaa ymmärtää, on olemassa tilanteita, joissa moraalinen agentti voi olla vastuussa ilman tietoisuutta moraalisesti relevanteista tilanteeseen liittyvistä seikoista. Sherin saamissa vastauksissa erityishuomiota kiinnitetään Michael J. Zimmermannin ja Angela Smithin kriittisille arvioille, Sherin implikoitua heidän joiksikin aikalaistavapuoikseen. Nykykeskustelun pääpiirteet kuvataan myös laajemmin, esitellen viisi siinä pääasiallisesti vaikuttavaa kantaa.

Vaikkakin Sherin argumentti on kiehtova ja todennäköisesti osittain moniin intuitiivisesti vetoava, esittelen huomattavia varauksia sitä kohtaan. Keskustelun nykytilassa on paljon erimielisyyttä koskien sitä, millainen painoarvo tiedolla tulisi ymmärtää olevan moraalisisessa vastuussa. Sherin esimerkkitalanteet, joissa hänen mukaansa nähdään intuitiivisesti vastuuta ilman tietoisuutta, huomataan olevan ongelmallisempia kuin mitä hän esittää.

Sen sijaan, että koettaisin vain sovittaa yhteen mahdollisesti yhteensovittamattomia ja eripuraisia intuitioitamme, ehdotan empiirisesti informoitua pragmaattista lähestymistapaa kysymykseen tiedon roolista moraalisisessa vastuussa. Esitän originaalin argumentin sille, miksi kohdeagentin tietoisuuteen ja tietotasoon on tärkeä kiinnittää huomiota arvioidessamme häntä ja kommunikoidessamme hänen kanssaan. Argumentti vihjaa, että vähintäänkin joissain tilanteissa tai aiheissa pätee, että jos tietoehtoa ei kunnioiteta tavalla, joka vaikuttaisi olevan yhteensopimaton Sherin selonteon kanssa, tällöin päädytään sekä moraaliseen paradoksiin että epäpragmaattisiin tuloksiin julkisessa viestinnässä. Tästä seuraavan normatiivisen, konsensientialistisen näkemyksen – *pragmaattisen näkemyksen* – huomataan eroavan vallitsevasta deskriptiivisestä, ansioperustaisesta keskustelusta koskien episteemistä ehtoa ja sen kodiksi ehdotetaan *yhdistelevää* tai *kosmopoliittista hyve-etiikkaa*.

Pragmaattisen näkemyksen teoreettisen kuvaamisen lisäksi myös soveltava vastaus tutkimuskysymykseen esitetään sekä sitä suunnilleen vastaava prosessikaavio. Pragmaattisen näkemyksen tämänhetkisen iteraatioon mukaan: moraalista tyrmistystä online ja offline voisi lievittää, julkista keskustelua parantaa, (moraalisesti relevantin) tiedon yhteiskunnallista välittymistä sujuvoittaa, rakentavampia mediakäytäntöjä edistää sekä poliittista polarisaatiota vähentää kultivoimalla esitettyä hyveellistä, pragmaattista heuristiikkaa tavoissamme arvioida toisiamme ja asettaa moraalista vastuuta. Kannustan pragmaattisen näkemyksen jatkokehittämiseen ja sen soveltamiseen paitsi yksityiselämässämme myös erityisesti tekoälyn ja sosiaalisen median alustojen kehittämisen ja sitä koskevan soveltavan etiikan kontekstissa.

Avainsanat: moraalinen vastuu, episteeminen ehto, moraalinen tyrmistys, sosiaalinen media, poliittinen polarisaatio, tiedekommunikaatio, metakognitio, moraalipsykologia, hyve-etiikka, pragmatismi

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

## Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
1.1	Moral Responsibility and the Contemporary Media Landscape of Moral Outrage	1
1.2	Progress and Climate Change	5
1.3	The Philosophical Interest: The Knowledge Requirement	9
1.4	Chapters	12
<b>2</b>	<b>HISTORICAL AND RECENT BACKGROUND</b>	<b>14</b>
2.1	Freedom and Knowledge: Aristotle's Two Necessary Conditions for Moral Responsibility	14
2.1.1	Voluntary, mixed, and non-voluntary actions	14
2.1.2	Involuntary actions	16
2.1.3	Culpable and non-culpable ignorance	18
2.1.4	Summary and Aristotle's influence	20
2.2	Outlines of Recent Discussion About Moral Responsibility	21
<b>3</b>	<b>THE SEARCHLIGHT VIEW – GEORGE SHER'S RECONSTRUCTION</b>	<b>25</b>
3.1	Sher's Definition of the Searchlight View	25
3.2	Indirect Evidence for the Appeal of the Searchlight View	26
3.3	Problems with the Searchlight View	28
3.3.1	The engaged and the detached perspective	29
3.3.2	Intuitions and practices	31
3.4	Imaginative Reconstruction of Failures to Justify the Searchlight View	34
3.4.1	Responsibility and practical reason	35
3.4.2	Responsibility and the Kantian Principle	39
<b>4</b>	<b>SHER'S ACCOUNT OF RESPONSIBILITY'S EPISTEMIC CONDITION</b>	<b>49</b>
4.1	Partial Account of Responsibility's Epistemic Condition	49
4.1.1	Application in Sher's nine example cases	51
4.1.2	Applicable standard	53
4.1.3	The boundaries of the responsible self	59
4.1.4	Summary of Sher's extrapolation of PEC	69
4.2	Full Account of Responsibility's Epistemic Condition	72
4.3	The Remaining Conundrum of the Voluntariness Condition	75
<b>5</b>	<b>RESPONSES TO SHER</b>	<b>78</b>
5.1	Michael J. Zimmerman's Argument for a Qualified Searchlight View	78
5.1.1	Examination of the views Sher rejects	79
5.1.2	Failure of common intuitive judgments: The origination thesis	81
5.1.3	Different kinds of agent evaluability?	83
5.1.4	Conclusion	85
5.2	Angela Smith's Attributionism	86
5.3	Remarks in Reviews: Psychological Limits, Doubts, and Controversies about Control	88
5.4	Outlines of Alternative Accounts of the Epistemic Condition	92
5.4.1	The role of awareness	93
5.4.2	The orthodoxy	96

5.4.3	Internalism, revisionism, and akrasia .....	97
5.4.4	Four answers to revisionism .....	99
5.4.5	Guidelines for reading chapters 6 and 7 .....	102
<b>6</b>	<b>A FURTHER RESPONSE: THE PRAGMATIC VIEW .....</b>	<b>104</b>
6.1	Sher's Example Cases and the Problem of Intuition .....	104
6.1.1	Our intuitions in Sher's nine example cases .....	105
6.1.2	An interlude of doubt .....	112
6.1.3	The problem with intuition .....	114
6.2	On Why the Engaged Perspective Should Always Matter .....	120
6.2.1	Interlinkedness of the engaged and the detached perspectives .....	120
6.2.2	Being responsible versus taking responsibility .....	122
6.2.3	Conclusion and epistemic states of agents .....	123
6.3	Epistemic Considerations: The Case of Bob and Global Warming .....	124
6.3.1	Introducing Bob and two views to his responsibility .....	125
6.3.2	The first scientist .....	127
6.3.3	Varieties of epistemic foundations .....	130
6.4	A Pragmatic Argument: The Case of Jack and Global Warming .....	132
6.4.1	Intuitions and practices .....	133
6.4.2	Empirical evidence in the case of Jack .....	135
6.4.3	How to adopt the pragmatic view, and implications for other accounts .....	147
6.5	A Clarifying Step Back: A More Distant Vantage Point .....	158
6.5.1	Different types of moral responsibility? .....	158
6.5.2	The pragmatic view as a predominantly normative view .....	160
6.6	Locating the Normative Home Neighborhood .....	161
6.6.1	Deep pragmatism and the tragedy of commonsense morality .....	161
6.6.2	Eudaimonia and naturalized Buddhism .....	163
6.6.3	Conclusion, application to social media design, and final caveats .....	167
6.7	Summary and Definition of the Pragmatic View .....	169
<b>7</b>	<b>HOW SHOULD WE THINK ABOUT THE KNOWLEDGE REQUIREMENT OF MORAL RESPONSIBILITY? .....</b>	<b>174</b>
7.1	Five Possible Scenarios and Our Probability of Accurate Agent Evaluation .....	175
7.2	Mitigating the outrage .....	179
7.2.1	The pragmatic heuristic .....	179
7.2.2	Some answers to possible perceived complications .....	184
7.3	Other Aspects of Virtuous Communication .....	186
<b>8</b>	<b>CONCLUSION .....</b>	<b>188</b>
	NOTES .....	193
	REFERENCES .....	237
	APPENDICES .....	271
	Appendix 1: Table of Two Intuitive Interpretations of Sher's Example Cases .....	271
	Appendix 2: Flowchart of the Pragmatic Heuristic for Agent Evaluation .....	272

# 1 INTRODUCTION

The themes discussed in this thesis cross several academic disciplines. Namely, those of moral philosophy, epistemology, moral psychology, Internet studies, and science communication. The primary focus, however, is on moral philosophy. As the secondary elements relate to and motivate the primary focus, the introduction likewise contains background information related to all of them. Much of the motivating research, statistics, and examples that are cited refer to English-speaking so-called Western nations, in many cases particularly the United States of America. Thus, the empirical material is admittedly mostly – though not entirely – WEIRD; that is, based on research on Western, Educated, Industrialized, Rich, and Democratic populations who, in the global scheme of things, are a minority and hence not necessarily representative of the general human population (Henrich, Heine, & Norenzayan 2010). However, the covered themes are much broader on the whole, as the topics of discussion have a lot to do with human thinking and interaction, and us coming together via the global network of the Internet.

Firstly, I provide a basic outline of the concept of moral responsibility, along with a dense summary of recent empirical evidence concerning polarizing moral outrage in our contemporary media landscape, especially online (1.1). Secondly, I illustrate how these problems also relate to failures of communication found within discussions of important current scientific topics like climate change (1.2). Thirdly, I introduce the primary philosophical interest of this thesis – the epistemic condition – that I see to be important in trying to better understand our current failures of communication (1.3). And, finally, I outline the structure of the chapters (1.4).

## 1.1 Moral Responsibility and the Contemporary Media Landscape of Moral Outrage

In our everyday lives, we often encounter situations where we feel a *moral norm* has been broken.<sup>1</sup> Especially nowadays, when we are in the virtual presence with a plethora of viewpoints practically 24/7, our exposure to these perceived *moral violations* has become more common than ever. Moral norms may roughly be characterized as rules, principles, dispositions, and character traits that are generally, at least in some group, agreed upon or felt to be worthy of upholding. Violations of these norms are deemed unwanted, wrong, even evil; and the agent – or group of agents – who is seen to be the author of the violations is often considered and held *negatively morally responsible*.<sup>2</sup> This is generally defined in terms of the perceived author being *blameworthy*, and thus worthy of *blame* directed at them.<sup>3</sup> However, moral responsibility may also be *positive*, in which case the perceived

author of a right or commendable act is considered *praiseworthy*, and thus worthy of *praise* directed at them. (see, e.g., Zimmerman 2009, 248; Shoemaker 2011, 604; Wieland 2017, 3; see also Zimmerman 2017a, 223–227.)

In recent years, on many news media outlets, but particularly on so-called social media – on platforms such as Facebook, Twitter, YouTube, Tumblr etc. – it has become increasingly common to express *moral outrage* (Crockett 2017a; Rost, Stahel, & Frey 2016; see also Huskey et al. 2018). Consequently, our various newsfeeds are filled with countless assignments of moral responsibility by various individual moral agents as well as groups of agents. Even though it is likely only a loud minority of people who actively generate and partake in the widely visible outrage (Cohn & Quealy 2019; DataReportal 2019), it does seem to be contagiously deepening social divides, tit-for-tat (Brady, Wills, Jost, Tucker, & Bavel 2017; Druckman, Levendusky, & McLain 2017; Pew Research Center 2016, 2017b; Suhay, Bello-Pardo, & Maurer 2017). To add fuel to the fire, people seem tend to exaggerate their differences, unwittingly, via generalizing many of the perceived moral violations to be strongly stereotypically descriptive of all individuals who are perceived to fall inside an out-group denominator (Graham, Nosek, & Haidt 2012; see also Brady et al. 2017). All of this may in part be contributed to the fact that all perceivable violations can be seen more easily by all possible interest groups via social media, compared to what could be seen on older mediums.<sup>4</sup> And similarly, the criticism concerning the perceived violations along with the assumed authors of that criticism, who are often considered morally responsible by opposing parties, are also ubiquitously visible. The resulting effects do not merely manifest as diminished quality of online deliberation, but also offline, as the polarized people online extend their effect to offline (Druckman et al. 2017; Hampton, Shin, & Lu 2017).<sup>5</sup>

Platforms such as Twitter, for example, facilitate mass movements of condemnation of – and between – moral agents and groups (Brady et al. 2017; Howell 2013; Suhay et al. 2017; see also Vosoughi, Roy, & Aral 2018). For instance, moral-emotional words on a tweet have been found to increase the likelihood of a retweet by approximately 20 % per word, especially within ideological group boundaries (Brady et al. 2017). Despite the possible virtues of some of these movements, they often happen without all the condemners – nor the ones being condemned, let alone the millions of potential spectators – being aware of all the facts of the situation that could, and often should, affect their judgments.<sup>6</sup> This can only be emphasized by online *echo chambers* that may largely insulate some groups of people from other viewpoints and arguments, forcing them to primarily hear and/or fortify their own one-sided, deeply biased views or interpretations (see, e.g., Brady et al. 2017; Bright

2018; Garimella, Morales, Gionis, & Mathioudakis 2018; Zollo et al. 2017; see also Frimer, Skitka, & Motyl 2017; Sunstein 2018; Villota & Yoo 2018; Visser & Mirabile 2004).<sup>7</sup> This sort of (self-)selective exposure has been found especially among groups of extreme positions, who tend to repel not only their counterparts but also moderate views (Bright 2018; Garimella et al. 2018; see also Friesen, Campbell, & Kay 2015; Druckman, Peterson, & Slothuus 2013; Pew Research Center 2017e, 2017f; Sunstein 2018). The reason for this repelling behavior may to some degree be explained by political extremism – both “left” and “right” – being more prone to believe conspiracy theories about other-minded political groups, in an aim to try to make sense of societal events (van Prooijen, Krouwel, & Pollet 2015; see also J. M. Miller et al. 2015). Tendency to believe political conspiracies is also connected to belief in simple political solutions to societal problems (van Prooijen et al. 2015) and to perceiving societal threats to control (van Prooijen & Acker 2015), among other things (for a systemic review and meta-analysis, see Goreis & Voracek 2019).

Pew Research Center has called these kinds of isolated groups “ideological silos”, while illustrating how in the United States ideological divide – that is, *political polarization* – is, by a wide margin, the widest it has been since 1994 when the surveys begun (2014, 2017b; see also 2018d). However, the divide has been traced to have begun its rise already in the 1980s, and particularly accelerating since 2000s, all the while motives for political participation have shifted from supporting a party to more so opposing one (Iyengar & Krupenkin 2018). At the same time, it seems many countries in Europe have also become increasingly divided, via rising populism (Norris & Inglehart 2019).<sup>8</sup> Perhaps not coincidentally, younger cohorts in many North American and Western European nations show increased dissatisfaction to democratic forms of government (Foa & Mounk 2016), though there is debate on what exactly this implies (cf. “Online Exchange” 2017). While echo chambers and other negative affordances of social media have likely played a role in the equation, perhaps the most consistent overall evidence for an explanation of the rising polarization has been gathered for the *cultural backlash theory*: silently risen socially-liberal attitudes, immigration, and economic grievances within the last decades have together triggered a conservative authoritarian reflex (Norris & Inglehart 2019). And this in turn can be seen to trigger a counter-backlash, overall resulting into a polarized state with (overly) highly attuned and dichotomous in-group–out-group sensitivities, particularly visible online (see also Suhay et al. 2017). In short, this can be framed as the tribalistic clash of globalist and nationalist ethos (Haidt 2016).<sup>9, 10</sup>

This kind of potential for mob mentality online has left a disproportionately negative impact on numerous individual lives for relatively minor offences, and in some cases has even caused fatal reactions to non-existent offences.<sup>11</sup> In some prominent cases, the impact can be traced to merely

having presented a sincere and nuanced argument – a call for further deliberation – that went against the mob’s often hearsay-based dichotomous assumptions that more so relied on uncharitable interpretations, common beliefs, and emotive reactions than carefully considered facts, knowledge, and arguments.<sup>12</sup> Consequently, many people feel they cannot even express their *moderate* views, without being pressured or threatened by some mob on one extreme side *and/or* another (Garimella et al. 2018; Hoffmann & Lutz 2017; Matthes, Knoll, & von Sikorski 2018; see also note 5).

Quite tellingly, false news has been found to spread significantly farther, faster, deeper, and more broadly on Twitter than true news, and especially in the case of political news (Vosoughi et al. 2018).<sup>13</sup> False news also tend to stir more emotionally disgusted replies (Vosoughi et al. 2018), indicative of generating more polarizing moral outrage (Brady et al. 2017; Suhay et al. 2017). At the same time, mere single exposure to false news has been found to increase subsequent perceptions of accuracy – i.e., producing *illusory truth effect* – and even if the news is explicitly labeled as contested by fact-checkers or is inconsistent with the reader’s political ideology (Fazio, Rand, & Pennycook 2019; Pennycook, Cannon, & Rand 2018). Furthermore, warnings of content having been disputed by fact-checkers may even backfire via increasing subsequent perceived accuracy of stories without warnings – i.e., producing *implied truth effect* – though this may be prevented by also attaching verifications to true news (Pennycook, Bear, Collins, & Rand 2019). Taken together, the above-mentioned are not good signs at a time when public trust in the traditional press is in many places considerably low or declining (Hanitzsch et al. 2018; Pew Research Center 2018a), and when false news online may soon be further accompanied with easy to accomplish AI assisted face- and voice-swapping in videos (see Gholipour 2017; Kim et al. 2018).

Unfortunately, experts seem equally split on whether polarizing misinformation and disinformation online can be mitigated (Iyengar & Massey 2019; Pew Research Center 2017c). Critical thinking seems to be a considerably scarce resource overall, but especially when it comes to believing what we read or see in favor of a perceived in-group or a related ideology (see Allcott & Gentzkow 2017; Čavojová, Šrol, & Adamus 2018; Frimer et al. 2017). Moreover, the problem doesn’t seem to be mitigated by people’s science literacy or numeracy: the better people’s science literacy and numeracy, the more able *and likely* they seem to be in interpreting words and data to falsely conform with their political outlook (Kahan 2017a; Kahan et al. 2012; Kahan, Peters, Dawson, & Slovic 2017b; see also Ballarini & Sloman 2017; Drummond & Fischhoff 2017; Joslyn & Haider-Markel 2014; Kahan & Peters 2017).

However, propensity for *analytic thinking* – measured via the proxy of performance in Cognitive Reflection Test (CRT) – may attenuate susceptibility to partisan false news, regardless of

one's ideology (Bronstein, Pennycook, Bear, Rand, & Cannon 2019; Pennycook & Rand 2019b, 2019c).<sup>14</sup> When and possibly only when combined with *epistemic rationality* (i.e., motivation to pursue and hold beliefs that accurately describe the world), general cognitive ability may also attenuate susceptibility to conspiracy beliefs as well as to paranormal beliefs (Ståhl & van Prooijen 2018). Additionally, even though more educated people can hold more polarized beliefs on controversial science topics (Drummond & Fischhoff 2017), education may still reduce belief in conspiracy theories insofar as it facilitates analytic thinking skills (or general cognitive ability and epistemic rationality), the insight that societal problems often have no simple solutions, a sense of control, and a sense that one is a valued member of society (van Prooijen 2017). Further, algorithmically utilized crowdsourced news source evaluation might be a surprisingly effective yet underutilized method to rank source quality, when it comes to familiar sources (Epstein, Pennycook, Rand 2019; Pennycook & Rand 2019a). Also, it would seem that cultivating *intergroup trust* would be an especially important project to develop when wanting to mitigate polarization (J. M. Miller et al. 2015; Nguyen 2018b). Still, even if susceptibility to partisan false news and conspiratorial thinking could be attenuated, and trust cultivated, partisanship might always remain a contextually destabilizing force in society, as we humans appear highly prone to thinking in terms of rightminded in-groups and wrongminded out-groups.<sup>15</sup>

All things considered, our ability to sincerely and civilly communicate with our fellow human beings, and openly deliberate about our ideas and future trajectories, seems to be seriously compromised in many ways. While the (near-)global reach of social media clearly has positive affordances for deliberation and distributing responsibility *in theory*, it seems we are not doing well at all in finding them *in practice*. Instead, we seem to be in many ways contributing to the atmosphere of moral outrage – regardless whether we are individually properly aware of our role in the equation.<sup>16</sup>

## 1.2 Progress and Climate Change

Living within the stormy shadow that our addictive social media landscape is habitually casting upon us, it is often easy to forget that *a lot* of things have gotten much better in the world and are still doing so. For example, on a global scale, in the past centuries and decades, there have been dramatic reductions in poverty, deadly diseases, and childhood deaths, along with substantial increase in nature conservation efforts, access to education, and basic literacy (Rosling, Rosling, & Rönnlund 2018, 51–64). And these kinds of trends are projected to continue.<sup>17</sup> I think this is a very important and therapeutic point to emphasize, especially in the current social climate. These accomplishments

should not be forgotten nor downplayed, as they may serve as an encouraging and empowering reminder that we *can* accomplish positive effects in the world.

At the same time, if we want to hold onto and continue on the path of our accomplishments, there does remain some substantially big problems and global catastrophic risk scenarios that we should collectively learn to pay better attention to. These include scenarios concerning, for example, a global pandemic, financial collapse, nuclear or otherwise largescale war, climate change, major impact event, major extinction event, disruptive technology (e.g., poorly controlled artificial intelligence(s) or poor utilization of biotechnology), or an unknown risk (Bostrom & Ćirković 2008; Rosling et al. 2018, 237–240; WEF 2018; WHO 2019; WWF 2018). To add, there is also the still lingering problem of extreme poverty (Rosling et al. 2018, 240–241; World Bank 2018). Unfortunately, some of these risks seem to have arisen or increased in tandem with our recent successes, insofar as those successes have either directly or indirectly relied on burning fossil fuels or contributing to other environmentally harmful or otherwise unsustainable or wasteful activities (see WWF 2018; see also IPCC 2014, 2018; UN 2015, 2016). The static noise of polarization on social media – while truly a disconcerting problem in itself – is most often doing nothing but force-feeding a stream of relatively unimportant pseudo-events, distracting and dividing us amidst bigger issues that should unite us all under a common cause.

In addition to the widely worrying state of our current media landscape and the polarized atmosphere within, I am particularly interested in science communication. In this context, especially climate science communication. Anthropogenic global warming (abbr. AGW) seems to be one of the biggest obstacles that stands in our way of continuing the path of healthy progress, and it is already affecting many areas of our lives (see IPCC 2014, 2–31; 2018; NASA 2019; Oreskes 2018; Raworth 2017; Ripple et al. 2017). Therefore, it would be in the best interests of everyone, if we could find better ways of communicating with each other – so that we would be able to make well-deliberated and informed political decisions in general, not least in terms of mitigating and adapting to global warming.

Unfortunately, our communication about science is not immune to similar challenges that are present more widely on social media. Arguably, it is infected by the same pandemic virus. For example, insofar as belief in AGW demographically best correlates with political affiliation, it is far from ideal that our newsfeeds are being filtered through our self-constructed echo chambers (see Hornsey, Harris, Bain, & Fielding 2016; Iyengar & Massey 2019; Williams, McMurray, Kurz, & Lambert 2015; see also Bolin & Hamilton 2018; Hmielowski, Feldman, Myers, Leiserowitz, &

Maibach 2014; Oreskes & Conway 2008; Vincent 2018).<sup>18</sup> Furthermore, moral outrage may in some circumstances be directed at third-party targets the more we feel our own moral character is in question, whether personally or via in-group association, consequently alleviating our guilt and restoring our moral identity – which all seems far from ideal in terms of communication and might only cyclically catalyze partisan outrage (Rothschild & Keefer 2017; Täuber & van Zomeren 2013). A distinct division in average beliefs relating to AGW can especially be seen between Democrats and Republicans in the United States (Pew Research Center 2015), or parts of the United States (Mildenberger, Marlon, Howe, & Leiserowitz 2015).<sup>19</sup> Yet, it seems that political deliberation and exchange of views online happens mostly near the political center, while those who hold extreme ideologies become effectively separated or alienated both from people of other viewpoints and even from people who hold more moderate versions of their viewpoint (Boutyline & Willer 2017; Bright 2018; Frimer et al. 2017; see also Pew Research Center 2017b, 2017f, 2018d). This effect is possibly explained by issues becoming more dogmatized and moralized at the extremes (Brady et al. 2017; Rollwage, Dolan, & Fleming 2018).<sup>20</sup> If we can't even communicate across ideological lines – and instead build social walls and see the “moral other” almost as an inherently irrational enemy – how are we supposed to successfully convey *any* important knowledge to each other?

Many studies have indicated that our political or religious group identity, along with associated moral values, is the guiding factor in our beliefs, and that science literacy and numeracy has little effect on the matter (for a meta-analysis relating to climate change, see Hornsey et al. 2016; Rutjens et al. 2018; see also Cohen 2003; Graham et al. 2009; Kahan et al. 2012; Pew Research Center 2017d). More specifically, this indication has been made on publicly controversial topics like AGW and vaccines in contemporary North America. Denying AGW has been found to be best predicted by political conservatism, while opposing vaccinations has been found to be best predicted by moral purity concerns and religious identity, and only the vaccine attitudes being statistically related to scientific literacy (Rutjens et al. 2018).<sup>21</sup> Religiosity, and particularly religious conservatism, has been found to consistently display low trust in science and unwillingness to support science (Rutjens et al. 2018). The connection of belief to ideological identity concerning other specific topics, however, varies. For example, opposition to GMOs has been connected with distrust of science and, indeed, low level of scientific literacy, while *not* being driven by political or religious identity (Drummond & Fischhoff 2017; Rutjens et al. 2018; see also Fernbach, Light, Scott, Inbar, & Rozin 2019). Still, considering the evidence regarding AGW and vaccines, the chance of us communicating important knowledge to each other may indeed seem bleak. It has even been indicated that education – including enhancement

of science literacy – may *increase* polarization on issues linked to political or religious identity (Drummond & Fischhoff 2017; Lewandowsky & Oberauer 2016).

However, the research examining the role of science literacy has, thus far, largely been based on items – usually less than ten – that measure very basic scientific knowledge.<sup>22</sup> Most often, the research has not measured the domain-specific scientific knowledge. In fact, it seems that only recently there has emerged research that has done just that. And the results are encouraging: communication of domain-specific scientific knowledge *may well* affect people’s beliefs across groups (Guy, Kashima, Walker, & O’Neill 2014; Hornsey et al. 2018; Lewandowsky & Oberauer 2016; Ranney & Clark 2016; Shi, Visschers, Siegrist, & Arvai 2016; Weisberg, Landrum, Metz, & Weisberg 2018; see also Milfont, Wilson, & Sibley 2017), and especially via group-specifically targeted communicators who frame the matter in an efficient way (Feinberg & Willer 2013, 2015; Hornsey & Fielding 2017; Täuber, van Zomeren, & Kutlaca 2015; Voelkel & Feinberg 2017; Wolsko, Ariceaga, & Seiden 2016; see also Albarracín & Shavitt 2018; Druckman & McGrath 2019; Kahan et al. 2011; Kahan, Landrum, Carpenter, Helft, & Jamieson 2017a; Farrell, McConnell, & Brulle 2019; Schuldt, Konrath, & Schwarz 2011).<sup>23</sup> There are also encouraging results on how communicating scientific agreement may neutralize motivated reasoning and thus politicization of facts, via correcting people’s perception of the scientific norm (Ding, Maibach, Zhao, Roser-Renouf, & Leiserowitz 2011; Lewandowsky, Gignac, & Vaughan 2013; Linden, Leiserowitz, & Maibach 2017; Ranney & Clark 2016). For example, merely communicating that “97 % of climate scientists have concluded that human-caused global warming is happening” seems to be well worthwhile (Cook et al. 2016; Linden et al. 2017).<sup>24</sup>

Moreover, if communicators could raise *science curiosity* in people – via inducing its emotional signatures of wonder and awe – it might counteract politically biased information processing (Kahan et al. 2017a).<sup>25</sup> In all communication, *civility* ought to also be upheld, for to do otherwise may both hinder the perceived strength of the argument and exacerbate polarization (Popan, Coursey, Acosta, & Kenworthy 2019; Sawaoka & Monin 2018).<sup>26</sup> To foster this all, *intellectual humility* appears to be connected to many beneficial prosocial characteristics, and might thus be something to promote as well, in an effort to facilitate productive intergroup communication. The prosocial characteristics connected with intellectual humility include, for example, curiosity, intrinsic motivation to learn, possession of more knowledge, better recognition of limits in one’s knowledge, openness during disagreement, possibly willingness to admit wrongness, and other features like less social vigilantism that could be seen to support civility (Fetterman, Curtis, Carre, & Sassenberg 2019; Krumrei-Mancuso, Haggard, LaBouff, & Rowatt 2019; Porter & Schumann 2018).

It has been noted in a review article that the evidence concerning the political divide on AGW and other topics can be read to suggest either that people are directionally motivated in their reasoning and thus reject any information that contradicts their beliefs, *or* that they are accuracy motivated but disagree on what counts as credible evidence (Druckman & McGrath 2019; see also Kunda 1990). These two explanations can vary individually and topically, and – I would think – they can be balanced in various ways situationally. Further, it would seem accuracy motivated people can disagree on who counts as a reliable testifier to any evidence. Regardless of what the aggregate balance of these explanations are on any given topic, researchers are beginning to find more fruitful ways of science communication for different cases (Druckman & McGrath 2019).

Alas, we are nowhere near to doing the best we could in conveying various knowledge to each other, especially considering the polarized atmosphere of moral outrage displayed on social media. Recently, due to us – or a highly visible portion of us – continuously bickering about relatively minor issues online, neuroscientist Molly Crockett has raised initial concerns about either possible outrage fatigue or sensitization, both of which can lower our capability to tackle problems that really ought to be tackled (2017b, 3:18–4:59; see also Brady & Crockett 2019). In short: we seem to be very keen on pointing fingers, dividing ourselves, and shouting at each other, rather than trying to sincerely understand one another and engage in truly important *dialogue* – and all of this is likely facilitated by the prevailing social media networks we have endowed ourselves to be governed with (Crockett 2017a; T. Harris 2017).

### 1.3 The Philosophical Interest: The Knowledge Requirement

The current day phenomena of polarized moral outrage and related challenges in science communication have motivated me to try to understand a bit better where we are going wrong in our interpersonal communication. I see the way in which negative moral responsibility seems to be taken for granted in some implicit form in the public sphere, and the consequent criticism dispensed in a very un-meditated fashion, to be signs that we ought to think more carefully about the matter and our relevant ways of thinking and acting. Moral responsibility likely has an important role to play in all stable societies as well as on the global Internet, and it seems to be an important feature of our social cognition, but our current patterns of distributing responsibility seem curiously polarizing and unproductive.

In this context, my philosophical interest lies in moral responsibility and the requirements for holding an agent morally responsible. Specifically, I am interested in the *knowledge requirement*, which is one of the two classic necessary requirements for moral responsibility, first described by ancient Greek philosopher Aristotle. Different contemporary thinkers have also referred to the requirement as the cognitive or mental condition, and the current orthodoxy is to talk about the *epistemic condition*. The other classic requirement is *freedom requirement*, also known as *or* sometimes divided into control requirement and voluntariness condition, and sometimes also called the metaphysical condition, but my primary interest lies in the knowledge requirement.<sup>27</sup> I see it to be in a key position in terms of understanding where our communication may be going wrong. Ultimately, in this wide context, the question I am interested in is this: *how should we think about the knowledge requirement of moral responsibility?*<sup>28</sup>

As there is some variation in the related terminology of different writers, it should be noted that I will be using the terms “freedom requirement”, “control requirement”, “control condition”, and “voluntariness condition” interchangeably, unless otherwise indicated. I will also use both terms “knowledge requirement” and “epistemic condition” interchangeably. Within the field, the precise difference between the knowledge and epistemic variations in most contexts seems unclear if not nonexistent, but it appears that the shift from “knowledge” to “epistemic” more broadly relates to a shift from an emphasis of knowledge to the emphasis on the possibility that the required mental state may not pertain to knowledge but rather to reasonable or justified belief, true belief, or simply belief, depending on the theory (Rudy-Hiller 2018, sect. 1.2; sect. 5.4.1.2). Thus, it may be understood that “epistemic condition”, more largely, is an umbrella of theories that contain any specifically knowledge-related formulations as well as other possible epistemic formulations of the condition for moral responsibility. However, just because I use both terms, it should not be deduced that I view the required mental state necessarily pertaining to knowledge proper.

I do utilize the term “knowledge” a fair amount, but quite loosely. Generally, it is used with a broad, intuitive, context-specific meaning. In many contexts, the focus is on scientific or propositional knowledge (“know-that”), but also often on procedural knowledge (“know-how”) and knowledge by acquaintance (“know-of”).<sup>29</sup> In formulating my own views, in section 6.1.1.1n107, I also more specifically define *metacognitive knowledge*, as it seems to be a less known form of knowledge and will at that point appear to be a necessary addition to the repertoire (it is recommended that one reads note 107 at that point, especially if not familiar with the concept). I also define a related concept of an *epistemic state* (in sect. 6.2.3). Thus, my own views concerning the kind of mental state involved appear to involve the question of what (meta)knowledge, or epistemic state, in a very general, context

specific sense an agent ought to have cognitively obtained (and retained) for them to be responsible for an act they are being evaluated of in the context they are being evaluated in (e.g., should the agent be aware of the acts consequences, moral significance, etc.). This terminology and the view behind them ought to become clear in the later chapters (6 and 7).

It should also be emphasized that whenever I'm referring to "responsibility", I'm referring to *negative moral responsibility*, unless otherwise specified, as that is my primary interest. In fact, the only sections that deal with positive moral responsibility are section 4.3, to complete George Sher's examination, and briefly section 6.7 to provide my own condensed thoughts, but otherwise the focus is on the negative side.<sup>30</sup>

I should yet like to emphasize that all assertions and claims made in this thesis should be taken to concern impressions or appearances or seemings. That is, everything I assert or claim appears for me to be the case, but my appearances may be mistaken (and/or those of my sources). As scientific epistemology strongly appears for me to provide the most trustworthy empirical understanding – for example, in terms of it providing empirical, probabilistic predictability and testability of hypotheses and theories – I take it to be the best available guide to the world we live in. Thus, insofar as someone has different appearances to mine, science ought to be the referee.<sup>31</sup> Insofar as scientific knowledge cannot settle some differences – for example, due to it being insufficient for the moment, or lacking in ability on some area – then all we have are arguments and appeals to our sentiments that ought to be approached with intellectual and moral humility, honesty, and charitability. If at some point I appear to be expressing too strong a certainty towards some assertion or claim, it ought to be taken as a product of being born into a space and time where dogmatic (interpretation of) language is pervasive. *Scientific skepticism* is undertheorized – almost to the point of being merely tacit – but it may approximately be defined as an epistemological attitude of provisionally accepting substantial empirical claims as knowledge only insofar as those claims survive rigorous critical evaluation, often with reference to our current (provisional, probabilistic) scientific understanding of reality or to methods utilized in science (see Sagan 1996; Novella 2018; see also Boudry & Pigliucci 2017; Hansson 2017; Kaufman & Kaufman 2018; Pigliucci & Boudry 2013). It is not always easy to follow – if even possible in a more overarching philosophically skeptical sense – but insofar as I try to follow some epistemic guideline, that is largely it. As for more of a moral guideline, that I am still searching, though I think this thesis comes closer to finding one than I've ever been before.

The stance we take on the role of knowledge in assigning moral responsibility has many implications on what we should consider ideal public communication – or, more specifically, ideal conducts of the

press in general, conducts of platform providers, and conducts of ourselves on social media and elsewhere. The implications are that much more important in the current atmosphere of increasing political polarization, while important knowledge-based decision should be made on many issues, perhaps most notably in relation to climate change. Of course, the schedule, efficacy and overall quality of those decisions are, at least to a significant degree, contingent on what the policy makers, and people who vote for them, know – or believe (to know) – about relevant matters. It is, therefore, of great importance to find out how we should think about knowledge in relation to responsibility.

## 1.4 Chapters

The chapters are largely in chronological order, both in terms of my steps of examination and the historical timeline of the views presented.<sup>32</sup>

In chapter 2, I present Aristotle’s original formulation of the two necessary conditions for moral responsibility. Additionally, I provide an outline of recent discussion, mainly *before* the epistemic condition started to gather attention.

In chapter 3, I move on to a key source of my examination: American moral and political philosopher George Sher’s book *Who Knew? Responsibility Without Awareness* (2009). Sher’s relatively recent book seems to be one of the most prolific, albeit rare, critical examinations of how we have understood and how we should (descriptively) understand the epistemic condition of moral responsibility. It is a useful starting ground also as it has inspired some decent amount of discussion about this often-neglected aspect of moral responsibility; the first edited collection of articles having only been published very recently (see Robichaud & Wieland 2017), and even more recently the first entry to the *Stanford Encyclopedia of Philosophy* (see Rudy-Hiller 2018). I present Sher’s arguments for why what he calls *the searchlight view* – which, roughly, sees that an agent cannot be morally responsible if they lack some relevant awareness – is an insufficient way to understand the epistemic condition.

In chapter 4, I present Sher’s proposal for an expanded alternative to the searchlight view, something that he calls the *full account of responsibility’s epistemic condition* (FEC) – which, roughly, sees that in some scenarios an agent can be responsible without awareness.

In chapter 5, I move on to critically examine Sher’s arguments. I introduce some reservations and counterarguments presented by some philosophers who have commented on Sher’s account, including Michael J. Zimmerman’s response to Sher and his argument for a *qualified* searchlight view

and the origination thesis (or revisionism), and Angela Smith's attributionism (or answerability). Furthermore, I provide an outline of the more overall discussion about the epistemic condition *after* it started to gather attention.

In chapter 6, I present my own thoughts about Sher's account and more generally about the role of knowledge in assigning moral responsibility. I make a case for why it would be important for the epistemic condition to get more attention in our discourse, public and otherwise. Specifically, in sections 6.3 and 6.4, by utilizing anthropogenic global warming as an example, I present a pragmatic argument for why Sher's account seems inferior to what I call *the pragmatic view*. Some essential clarifications, however, are expressed in section 6.5, and the importance of the pragmatic view is emphasized in section 6.6. I underline how I see the pragmatic view to be important in approaching the topical question of how we could better approach one another and cooperate in a world that is rapidly becoming globally connected. The challenge is that our moral perspectives seem to be largely at odds, but I do not think that challenge is insurmountable. Relatedly, I connect the pragmatic view to normative, scientifically informed, fusion or cosmopolitan virtue ethics – and especially naturalized Buddhist *eudaimonia*.

While chapter 6 can be read as a theoretical answer, in chapter 7, based on the pragmatic view, I formulate my applied answer to the research question of how we should think about the knowledge requirement of moral responsibility. Specifically, I argue that we ought to cultivate following a pragmatic, virtuous heuristic when evaluating perceived transgressors, whether online or offline but nowadays particularly online.

Final remarks and a summarizing outline of the thesis can be found in the concluding chapter 8. Overall, the thesis can be read as an analysis of the epistemic condition and as an answer to Sher, but also as a multidisciplinary exploration towards mitigating the problems outlined in sections 1.1 and 1.2. It is worth mentioning that many of the longer notes accompanying this thesis add a lot especially to the latter endeavor (particularly in chapters 6 and 7), and are thus recommended especially for those who are interested in that aspect. I would also hasten to add that even though the examination of Sher's position is heavily philosophical, the later chapters become much more practical, and also more interested in relevant psychology. Thus, some readers may opt to skip some sections (the pragmatic meat of the matter, so to speak, can be found in chapters 6 and 7).

## 2 HISTORICAL AND RECENT BACKGROUND

To provide some background for the subsequent examinations, I outline two aspects of past discussion: the historical starting point of Aristotle's examination (2.1), and a basic characterization of the discussion roughly before the contemporary focus on the epistemic condition emerged (2.2). Later, in chapter 5, after having explicated Sher's account in chapters 3 and 4, I outline the recent discussion specifically in terms of how others have reacted to Sher's novel examination and what discussion has been generated about the epistemic condition more generally.

### 2.1 Freedom and Knowledge: Aristotle's Two Necessary Conditions for Moral Responsibility

In Book III of the *Nicomachean Ethics*, ancient Greek philosopher Aristotle (384–322 BCE) sets out to clarify the nature of virtue (in Greek: *aretê*), after first having introduced the subject matter in the two preceding books.<sup>33</sup> Specifically, our interest lies in chapter 1, where he divides actions into two categories: voluntary (*hekousion*, *hekôn*) and involuntary (*akousion*, *akôn*). The former are acts for which an agent is held responsible for and the latter are acts for which responsibility is pardoned. The distinction between these, however, is not very clear-cut, which leads Aristotle to expand from them to two other categories that reside somewhere between the two original ones: mixed and non-voluntary actions. Written approximately 350 BCE, these four distinctions contain the earliest known precursors for what are now widely considered as the requirements, or conditions, for moral responsibility (e.g., Rudy-Hiller 2018; Sher 2009, 3–4; Zimmerman 2017a, 219). The two necessary conditions – also usually taken to be jointly sufficient conditions – emerge from this division; specifically, via Aristotle's examination of the involuntary acts. In this section, I introduce Aristotle's four categories of action, and illustrate how the two conditions for moral responsibility derive from them. In the examination, I have cross-referenced translations by C. D. C. Reeve (2014), H. Rackham (1934), and W. D. Ross (1980), so all the references below refer to the aggregate process, while all the quotations and most terminology in the body text are from the Reeve translation.

#### 2.1.1 Voluntary, mixed, and non-voluntary actions

A *voluntary action*, in Aristotle's account, is an act that an individual performs willingly, usually by deliberate choice, to introduce changes in the world in order to achieve a certain goal – and is thus usually held morally responsible for. This is an action that has as its starting-point the agent themselves,

who – in contrast to involuntary action – is not acting under compulsion nor are they ignorant of particular facts of the situation, at the time of the act.<sup>34</sup> (III.1.1110a12–18; III.1.1111a22–24; see also III.2–5.) In general, Aristotle states that these are acts that are praised and blamed (III.1.1109b30–35).<sup>35</sup> For example, if I *freely and deliberately decide* to throw a rock towards a person, *intending* to bring about an injury in them, and then successfully *act* upon that decision, I am most definitely held morally responsible for injuring them and hence most likely blamed (assuming I have not been able to hide my doing the act, and assuming that injuring someone with a rock is against a significant moral norm or representing an action devoid of virtue).

*A mixed action* is an act that no one would willingly choose to perform *as it is* or *on its own*, but one that may be done reluctantly in some situations where all other options available are considered even worse. In other words, these are actions in situations where we are forced to choose between “two evils” or more. It is less clear whether one should be held responsible for mixed actions in different cases. One example that Aristotle mentions is throwing cargo overboard in a storm, when it *may* help secure the safety of the ship’s crew. It is not entirely clear whether this act would be commendable, or reproachable, or neither, or both in different respects. Or, as a more contemporary example: in the classic *trolley problem*, first variation formulated by British philosopher Philippa Foot (1967), where one decides whether to pull a lever and save five people but in doing so sacrificing one person, it is less clear whether one should be held responsible for letting the one person die, as doing so saves five and the choice is binary. Aristotle seems indecisive whether to consider these actions a subgroup of voluntary actions, responsibility thus remaining, as the starting-point of the act is within the agent; or as involuntary actions since no one would choose to perform such an act *as it is*, but the choice appears to be forced by the circumstances. He does say that these are *more like* voluntary acts. It could be described that Aristotle thinks responsibility in the case of mixed actions is determined on a case-by-case basis, based on whether the agent displays a virtuous character via their manner of ranking the “two evils” (see Campos 2013, 110–111, 117–119; Pakaluk 2005, 124–126, 128–129). For example, if I throw cargo overboard at the very first sign of a storm, I am likely acting in panic that is not commendable (and it may or may not be reproachable, as I am likely to be an inexperienced, unknowledgeable sailor). But if the decision is made in a situation that truly seems dire, with proper yet undoubtedly quick deliberation, it may very well be commendable. (III.1.1110a3–30.)

*A non-voluntary action* is an act done in ignorance of particular facts of the situation, but one that does not induce pain and regret in the individual after they later find out about their ignorance.<sup>36</sup> Conversely, if an individual does feel pain and regret after learning about their ignorance, they have

not acted non-voluntarily but involuntarily instead (I will get back to involuntary actions in more detail in the next section). It seems that someone who has acted non-voluntarily is responsible insofar as they have indeed broken an important moral norm (not for the act *per se*, but for their character which seems to display a disposition of not caring to prevent the kind of negative act in question). For example, we may imagine a busy night out with friends: If I accidentally mistake a drink on a table to be mine even though it is someone else's, and fully drink it, there is likely to be little pain and regret afterwards if I can easily offer to replace the drink after having learned of my error – and I am thus pardonable. Still, if I do not feel pain and regret even to the degree where I would offer to provide a replacement, responsibility seems warranted. However, if I only take a sip of the drink before noticing my error, it seems unclear if we should think much anything of the situation in terms of responsibility (regardless of whether I feel regret or not). In that case, it was just a *negligible* mishap. (III.1.1110b16–24; see also III.1.1110b10–12; Campos 2013, 105–109; Pakaluk 2005, 123–124, 127.)

As can be seen, the boundaries between voluntary, mixed, and non-voluntary acts are not always very clear, especially in terms of whether the agent displays bad character and is therefore considered responsible. Thus, all these categories of action could be discussed much further (see, e.g., Campos 2013; Pakaluk 2005). However, our primary interest, in this context, lies in the fourth category that is in direct opposite to voluntary actions.

### 2.1.2 Involuntary actions

*Involuntary actions* are acts that people – or what we now more precisely call *moral agents*<sup>37</sup> – are not usually considered responsible for (III.1.1109b30–35). When performing an involuntary action, an agent acts while either [A] being compelled to do something by *external force*, i.e. the act is compulsory, being out of their control or chosen without freedom, or [B] being *ignorant* of the particulars in which the action lies and with which it is concerned, i.e. the act being done due to ignorance of particular facts of the situation and thus producing both pity and sympathetic consideration or pardon, potentially making the agent not responsible (III.1.1109b36–1110a4; III.1.1110b1–8; III.1.1110b30–1111a2). However, Aristotle emphasizes that the act must also have caused pain and regret in the agent for him/her to be considered involuntary due to ignorance – meaning also, at least in the case of ignorance, that the final judgment of the act happens *a posteriori*, unlike in the case of voluntary acts which may be evaluated during the act (III.1.1111a16–21). Thus, either during or after [A], or after [B], if the agent realizes that had they not been ignorant of the

particular facts of the situation they would have acted differently, and if as a consequence of [B] they feel pain and regret, then the action is considered involuntary.<sup>38</sup> For example, even though not strictly moral responsibility, adaptations of these kinds of principles may be found in criminal law, in principles concerning legal responsibility: for instance, it is commonly considered that a person needs to achieve a certain age – implicating a certain state of cognitive and intellectual development (similar to freedom and non-ignorance) – before they can be held legally responsible; that is, the age of majority.<sup>39</sup>

In involuntary acts, in the case of external force that can compel an agent to do the act, the force is one that has as its starting-point something external to the agent – to which the agent “contributes nothing”. Aristotle himself uses the examples of wind or other human beings with control over the agent. (III.1.1110a1–4; III.1.1110b1–18.) Insofar as the agent has contributed to the act, for example having consciously yielded to ill persuasion, the act is no longer involuntary but instead voluntary or mixed, depending on further details.

Concerning involuntary actions in the case of ignorance, Aristotle explicates what he considers to be included in “the particulars in which the action lies and with which it is concerned”. These are the aspects of the situation that an agent can be ignorant of, in the moment of the act, and that seem to be encouraged to be evaluated *a posteriori* – in relation to the agent’s pain and regret – when deciding if an act was involuntary, and thus if it warrants a pardon from responsibility. According to Aristotle, when someone acts, we can ask “[1] who? [2] what? and [3] concerning what? or [4] in what? and sometimes [5] with what? (for example, with what instrument?), [6] for the sake of what? (for example, preservation), and [7] in what way? (for example, weakly or intensely).”<sup>40</sup> More explicitly, revealed more clearly via Aristotle’s examples, it seems these refer to the following type of questions, the answers to which the agent could be ignorant of during the act: [1] who is doing the act?; [2] what is the act s/he is doing?; [3] who or what is the subject of the act (that is being affected)? or [4] what are the (interpersonal) circumstances or location of the act? (*see previous footnote 40*); [5] what is the physical object the act is being done with?; [6] what is the effect of the act? (i.e., the agent can be ignorant in terms of the act not having the effect they intend, or it not having the lack of effect they intend); and [7] what is the proper intensity or other such qualitative aspect of the act? (i.e., the agent can be ignorant of the right way to perform the act). According to Aristotle, only a madman would be ignorant of all the particulars; or of [1], i.e. that *they* are doing the act. Of each other particular, [2]–[7], a mentally healthy agent can be ignorant. (III.1.1110b30–1111a20.)

Aristotle seems to put emphasis on [4] and [6], which he identifies as us having the most control over. In Reeve’s translation, these refer to the circumstances of the act and the effect of the

act. Reeve interprets Aristotle to mean that if we are ignorant about either of these, it will make our knowledge or ignorance of the others irrelevant, and we are automatically considered having acted involuntarily and thus redeemably (given that we show signs of pain and regret after our ignorance is cured). In all translations, Aristotle does seem to say that being ignorant of any of the other particulars suffice as well, however he emphasizes the two abovementioned. I should note that it might also be interpreted that ignorance of either [4] or [6] would result in an involuntary act, redeeming the agent of responsibility, but otherwise the responsibility would be evaluated on a case-by-case basis on an unspecified continuum of degrees of responsibility (and a virtuous person could perform the evaluation prudently). However, I think Aristotle is merely emphasizing [4] and [6] as the one's we have the most control over, but not disregarding the value of the other particulars as grounds for an involuntary act. (III.1.1111a15–20; see also Aristotle & Reeve 2014, 243n213.)

### 2.1.3 Culpable and non-culpable ignorance

One further nuance relating to actions in the case of ignorance can be found, though the corresponding part of Aristotle's writing seems especially obscure. There is thus some uncertainty here. With the aid of Pakaluk's interpretation, however, it may be emphasized that for ignorance to be qualified as involuntary, it needs to indeed concern the *non-moral* (or 'factual') particulars of the situation, as illustrated above, *not* ignorance of [i] moral particulars nor [ii] moral generalities. Furthermore, it may be interpreted that there are some exceptions to [i] in the sense that in some cases also non-moral ignorance may warrant responsibility. This is comparable to what is nowadays known as the distinction between *culpable* and *non-culpable ignorance*, where responsibility and blame are commonly viewed to be appropriate only for the former.<sup>41</sup> Respectively, Aristotle can be read to make the distinction by referring to the difference between acting "because of ignorance" (culpable) and acting "in ignorance" or "while ignorant" (non-culpable). (III.1.1110b24–1111a2; Pakaluk 2005, 126–127, 127n6).

It helps to remember that Aristotle's goal here is to see when we can draw an inference from an agent's action to their character, and thus to them lacking some character-related virtue; in other words, moral virtue (see note 33). This inference becomes possible whenever an agent seems to lack some moral virtue during the act (i.e., when ignorant of moral particulars), or whenever they seem to lack awareness of relevant basic concepts or principles (i.e., when ignorant of moral generalities). Following Pakaluk's example, in the case of ignorance of moral particulars: we may imagine a situation of a large transaction where it has previously become clear to the seller that the newly issued

bills tend to stick together. Let's then say that the seller does not pay careful attention to the transaction being fair, and it is later found out that some extra bills stuck and the seller unwittingly took more than agreed upon. In this situation, the seller may be considered culpably ignorant in the sense that they lacked the virtue of fairness (or "justice"), and thus displayed bad character. Further, in the case of ignorance of moral generalities: if the seller then keeps the excess money because they are not aware that a seller should only receive the amount fairly agreed upon, or because they do not realize that in business deals fairness is preferred over monetary gain, they would also be culpably ignorant. (III.1.1110b24–1111a2; Pakaluk 2005, 127–129.)

After his examination in chapter 1, Aristotle touches on culpable ignorance once more in chapter 5. It may be interpreted that in chapter 5 it is somewhat confusingly revealed – or at least more clearly revealed – that there are, after all, some instances where actions done in ignorance of some *non-moral* particulars may also count as culpable, voluntary actions. Aristotle seems to imply that we can also be responsible for our non-moral ignorance if we are responsible for that ignorance. The example Aristotle gives is of drunkenness: an agent may be responsible for their drunken actions even though the actions may be done in ignorance, insofar as the starting-point for the ignorance is in the agent (that is, in their character of willfully becoming intoxicated to the point of risking ignorance). Moreover, Aristotle seems to imply that we can be responsible for our non-moral ignorance if we break the law while ignorant of provisions of the law that we ought to have known and that could have been easily known by us. In general, it seems Aristotle views responsibility for non-moral ignorance to be appropriate for an act if it was done in easily (self-)curable ignorance that the agent *neglected* to cure *a priori*: that is, when they knew of their (potential) ignorance, had good reason to want to cure (or prevent) the ignorance, and when it would have been reasonably within their power to do so, yet they didn't do so. (III.5.1113b30–1114a31; see also III.1.1110b24–1111a2; Pakaluk 2005, 146–147.)

As an interpretive point, it may be read that this form of responsibility would, after all, come back to moral ignorance, as it appears to be the source of this kind of non-moral ignorance (in Aristotle's examples). That is, it appears that neglecting to cure a particular non-moral ignorance of which an agent is aware to be an important ignorance to cure, and that would be easy for them to cure, would imply the agent lacking some moral virtue due to having developed ineffectual habits. For example, if it is within our character to casually become intoxicated to the point of ignorance, it would appear that we would ultimately, in minimum, lack the moral virtue of temperance (III.10–12). Or, if it is within our character to neglect becoming informed about a law that would easily be in

our means to become informed of, it would appear that we would ultimately, in minimum, lack the moral virtue of justice (V).

Especially in scenarios of culpable ignorance, but possibly also in other blameworthy scenarios, I am prone to interpret Aristotle to think that moral responsibility, and associated blame, is an appropriate punishment for the agent at least in part in the sense that it would guide their character to a better direction. After all, we would all seem to be each other's environmental authorities that can mutually guide each other toward cultivating virtue or fail to do so. (see note 33; III.5.1113b20–30; X.9.) However, I cannot help but think that contemporary psychology and medicine would be more understanding and, as a result, reserved in some cases of what Aristotle would count as culpable ignorance. For example, our contemporary knowledge about alcoholism, addiction, and depression – including their possible external and biological causes (i.e., starting-point being outside of the agent, in some potentially significant sense) – could have had a noticeable effect on Aristotle's views, had he been aware of the relevant knowledge (cf. III.5.1113b30–1114a10).<sup>42</sup>

#### 2.1.4 Summary and Aristotle's influence

Overall, to summarize the most relevant part of Aristotle's account, the conditions for an *ignorant* agent – or at least for a non-morally ignorant agent – to be involuntary and hence pardonable from responsibility for an otherwise blameworthy act, three things must occur: [1] the agent must overcome the ignorance that they were under during the act, [2] the agent must experience (psychological) pain in consequence, after acquiring the knowledge that they were ignorant of during the act, and [3] the agent must regret their action. If only condition [1] is met, and the act was not externally forced, then the agent has acted in a mixed or *non-voluntary* manner, and thus it becomes less clear whether they are to be considered responsible. If all the conditions are met, then the act was *involuntary*, and hence pardonable.<sup>43</sup> (see especially sect. 2.1.1–2.1.2.)

There is a lot of ambiguity in Aristotle's account of responsibility, especially concerning mixed and non-voluntary actions, and how they relate to responsibility (cf. Campos 2013; Pakaluk 2005, 119–129). Furthermore, his list of the seven particulars in which the action lies and with which it is concerned could be discussed further, along with culpable (and non-culpable) ignorance. For example, relating to self-control, Aristotle would appear to think that an agent is responsible if they consume alcohol or drugs in a situation where they should retain their cognitive faculties, if they know the substances would compromise their character. Or, relating to culpable ignorance, it could be emphasized that Aristotle might think an agent is responsible when ignorant if they have not

utilized their acquired information-seeking skills to cure that ignorance, even though they had good reason to think they should, and their available means to utilize them were sufficient.

Regardless of the missing nuances, amidst his examination, Aristotle distinguished two types of circumstances that are required for an action to be (potentially) involuntary, and thus pardoning the agent from responsibility: [A] the act is performed under external force or compulsion, i.e. it is done out of the agent's control or without freedom; and [B] the act is done due to ignorance (of certain kind).

Ever since Aristotle, according to George Sher (2009, 3–4), philosophers have widely accepted that for a person to be morally responsible, two corresponding necessary conditions need to be met: the first relating to the will of the person (the act must be voluntary), and the second relating to the knowledge a person has (the act must not be done because of ignorance).<sup>44</sup> Initially, Sher calls the former *freedom requirement*, and the latter *knowledge requirement* of responsibility.<sup>45</sup> According to Sher, even though philosophers have discussed the freedom requirement rather thoroughly, discussion of the knowledge requirement has largely been neglected, and – among academic philosophers – taken for granted in an implicit form that he calls *the searchlight view*.<sup>46</sup> The general view among the philosophers who have started to take the knowledge requirement as the starting point of analysis for moral responsibility, as opposed to the more traditional analysis starting from the freedom requirement, is that the discussion is greatly affected whether the analysis starts from one or the another (Wieland 2017, 5). Furthermore, as the knowledge requirement may, in some cases, be sufficient to render an agent fully blameless all by itself, no matter if other conditions are met, it is important to find out when this might be (Wieland 2017, 5).

## 2.2 Outlines of Recent Discussion About Moral Responsibility

Aristotle's overall examination of responsibility is commonly seen to be imprecise in one especially notable way. Namely, it is not entirely clear what it means for it to be *appropriate* to direct praise or blame to an agent. Specifically, at least two interpretations are possible: [A] praise or blame is appropriate in the sense that the agent *deserves* such a response, due to their (voluntary, non-ignorant) action and/or traits of character, or [B] praise or blame is appropriate in the sense that such a reaction would likely bring about a desired consequence, namely an improvement in the agent's behavior and/or character. These two interpretations may be seen to correspond to two competing interpretations of the concept of moral responsibility: (1) the *merit-based view* (or desert-based view), which considers praise or blame as appropriate reaction toward an agent if and only if they merit, i.e.

‘deserve’, such a reaction via the nature of their act; versus (2) the *consequentialist view*, which considers praise or blame as appropriate reaction if and only if the reaction would likely lead to a desired change or reinforcement in the agent and/or their behavior. Respectively, the former is often characterized as a backward-looking view (paying attention to what the agent has done in the past), and the latter a forward-looking view (paying attention to the future effects to the agent, others, and/or society). In the last 50 years, the focus on the field has overwhelmingly been on offering alternative accounts of the merit-based view, and on questioning whether there is only one concept of moral responsibility. As hinted at the end of section 2.1.3, I am myself more attracted to the consequentialist view, though not in the sense that the appeal would depend on what the correct interpretation of Aristotle might be. (Caruso 2018, ch. 1; Eshleman 2016, ch. 1; Wieland 2017, 4; see also Eshleman 2016, note 7.)

The distinction between merit and consequence rose to prevalence approximately since the Stoics in the 3th century BCE, when the thesis of *causal determinism* began to play a central role in theorizing about moral responsibility. Causal determinism is the view that everything that happens or exists is caused by sufficient antecedent conditions, making it impossible for anything to happen or be other than it does or is.<sup>47</sup> Nowadays, especially influential version of causal determinism is *scientific determinism*, which identifies the relevant antecedent conditions as a combination of prior states of the universe and the laws of nature.<sup>48</sup> Relatedly, moral philosophers today may be classified into two types: (1) *incompatibilists* about determinism and moral responsibility, who maintain that if determinism is true, then there is nothing for which one can be morally responsible, and (2) *compatibilists*, who think that even though determinism is true, a person can still be morally responsible for some things. (Eshleman 2016, ch. 1.) Incompatibilists may further be divided into two main subcategories depending on how they react to the incompatibilist stance: (1a) *free will skeptics*, who argue that because determinism is true, free will is not; and (1b) *libertarians*, who argue that because free will is true, determinism is not, at least not in a universal form (Caruso 2018, ch. 2.1–2.2; Eshleman 2016, ch. 1; Talbert 2016, 16). As the libertarian view tends to be connected with dualistic religious commitments, those inclined to non-religious examinations tend to dismiss the libertarian view as thoroughly incoherent (e.g., S. Harris 2012; Dennett 2014). By and large, the discussion between these views has for a long time focused on discussing the freedom/control requirement, and whether free will/control is, in some sense, ‘compatible’ with determinism (see, e.g., Russell & Deery 2013; Talbert 2016, 14–29).<sup>49</sup>

Historically, these views have roughly corresponded with the merit-based view and the consequentialist view: as free will skeptics view that no one is in ultimate control of their actions and

no one can choose other than they do, merit-based blaming and praising make little sense (Caruso 2018), whereas libertarians have tended to advocate the merit-based view (Eshleman 2016, ch. 1). Compatibilists apparently tend to accept the consequentialist view (Eshleman 2016, ch. 1), though insofar as they hold others responsible in a backward-looking sense they would be sneaking in merit/desert (cf. Caruso & Dennett 2018).<sup>50</sup>

Next, I focus on introducing some general characteristics and distinctions in recent discussion about what constitutes someone as morally responsible. The contemporary positions may largely be read as merit-based views (and thus compatibilist or, more rarely, libertarian), but a consequentialist may also read them as hypotheses for what would provide desired cognitive or behavioral changes in the target-agent(s). The two most often utilized recent approaches have been characterized in the overlapping terms of responsibility as *accountability* and *attributability* (Shoemaker 2011; Watson 1996; see also Eshleman 2016; Talbert 2016).

Accountability theorists maintain that to be morally responsible is to be recognized as the author of an interpersonal violation, under some theory-specific moral standards that the author could reasonably have been expected to meet yet fail to meet, and thus as an apt candidate for *participant reactive attitudes* (e.g., Darwall 2006). Reactive attitudes, further, refer to the psychological disposition of people to react – via blame – to a perceived author’s perceived quality of will behind their perceived violation, which reaction may be suspended if the target-agent is excused (e.g., due to the act being an accident) or their behavior deemed justified (e.g., in the case of an emergency), or the agent is deemed psychologically/morally abnormal or undeveloped (see P. F. Strawson 1962/2013). (Caruso 2018, ch. 1; Eshleman 2016, ch. 2; Wieland 2017, 5.)

Attributability theorists, on the other hand, maintain, within some theory-specific limits, that to be responsible is for the violation to disclose an evaluative judgment, commitment, or character of the author (e.g., Sher 2009). Among these theories, reactive attitudes are not necessarily mentioned, but instead responsibility might simply involve fault or credit being attributable to the apparent character of the target agent. There can be overlap between accountability and attributability accounts, and they may be viewed as different types of responsibility, and thus not necessarily competing. More largely, there are disagreements about, for example, what constitutes excuse and what kind of reactive attitudes are warranted in different cases, or what kind of cases warrant fault or credit. (Caruso 2018, ch. 1; Eshleman 2016, ch. 2; Talbert 2016, 48–49; Wieland 2017, 4.)

Not all views fit within the two categories, however. For example, *answerability* theorists can be seen to try to unify the two via maintaining that to be responsible is for the violation to be connected to the author’s capacity for evaluative judgment in a way that demands of justification can be directed

at them (see, e.g., Scanlon 1998; A. M. Smith 2012).<sup>51</sup> Furthermore, morally skeptical views maintain that no conditions on being morally responsible can ever be met and thus no one is ever responsible (see, e.g., G. Strawson 1994/2013). Partly in attempts to either answer or work with the skeptical perspective, revisionists offer further views that reconstruct moral responsibility (see, e.g., Pereboom 2013, 2015; Scanlon 1998, 274–277). (see also Caruso 2018; Eshleman 2016, sect. 2.2.)

Moreover, it has been suggested that no single theory captures the fundamental criteria behind our assignments of moral responsibility as people seem to utilize different criteria in different kinds of cases (e.g., Knobe & Doris 2010). Whatever the philosophical constituents for moral responsibility may or may not be, it seems clear that people do widely blame each other, in various ways, for seeming to have broken various moral norms. For example, some relevant empirical evidence that fit together with accountability have been described in moral psychology (see Dill & Darwall 2014; Haidt 2012, 87–89). As also demonstrated in sections 1.1 and 1.2, there does seem to be a lot of (hastily) reactive attitudes going on, albeit they could potentially, *prima facie*, be interpreted via any of the abovementioned lines of theory. It is noteworthy, however, that accountability seems not to be compatible with free will skepticism as it appeals to backward-looking merit, but attributability and answerability can in some variations as they do not necessarily appeal to merit nor, it seems, to backward-looking praise nor blame (Caruso 2018, ch. 1; see also Caruso 2018, sect. 3.2; note 50 above).

As will become clear, despite the majority of contemporary positions appearing to examine the epistemic condition via the merit-based view, I am myself prone to examine and evaluate those positions from a consequentialist perspective, and eventually formulate a scientifically informed consequentialist view of the epistemic condition. This may be seen as being encouraged by the outrage on social media that our current habits are manifesting to the point of seriously poisoning public discourse (see sect. 1.1 & 1.2). The shift from merit to consequences will particularly take place in chapter 6, especially section 6.1.3. But before that, it helps to understand the particulars of the only recently emerged focus on the knowledge requirement or the epistemic condition. Chapters 3–5 focus on understanding and outlining the particulars of the relevant discussion, while chapters 6 and 7 focus on providing a justification for why a move to a consequentialist framework might be in order and what it might look like.

### 3 THE SEARCHLIGHT VIEW – GEORGE SHER’S RECONSTRUCTION

In his book *Who knew? Responsibility Without Awareness* (2009), George Sher takes the initiative to critically examine the knowledge requirement of moral responsibility. Largely via his own reconstruction, he discusses the prevalent view he sees modern academic philosophers to take in relation to the knowledge requirement and names it as *the searchlight view*. He then moves on to critically examine it, and to develop what he sees as a more viable alternative.

In this chapter, I present what the searchlight view is, as Sher sees and reconstructs it (3.1), examine what evidence Sher presents for his claim that it is implicitly prevalent among academic philosophers (3.2), what his objections to the searchlight view are (3.3), and what reconstructed lines of theory he considers failing as possible attempts of justifying the searchlight view (3.4).

#### 3.1 Sher’s Definition of the Searchlight View

Sher derives the name “searchlight view” from an analogy that he sees as descriptive of the prevalent position among academic philosophers, concerning the knowledge requirement of moral responsibility. A searchlight is a device we use to locate something in the dark, and the light emitted can be more or less narrow or wide, and more or less dim or bright. When trying to discover something with a narrow and dim searchlight, it can be very difficult, even impossible; and the chances of finding out what we are looking for can significantly be increased by using a bright and wide searchlight. The widely held view, according to Sher, is that the agent’s consciousness is analogous to a searchlight: it too can be more or less narrow, and more or less dim – and if the agent’s consciousness, because of its dimness, for example, is unable to attain some important possibility relating to doing or omitting a certain action, or relating to relevant outcomes, then this inability of the agent to focus the searchlight of their consciousness tells us that the agent is not – and likely should not be held – morally responsible. (Sher 2009, 5–6.)

According to Sher’s central interpretation of the searchlight view, an agent’s responsibility extends only as far as their awareness of what they are doing. An agent is then, under this view, only responsible for *acts* s/he consciously chooses to perform, *omissions* s/he consciously chooses to allow, and for *outcomes* s/he consciously chooses to bring about. In other words, an agent is only responsible if the relevant features of the act were “illuminated by the searchlight of his consciousness.” The focus in responsibility is solely on the conscious aspects of the agent and their

actions; solely on the acts, omissions, and outcomes that the agent is consciously aware of. On the other hand, the focus is on the agent's *searchlight control*: the ability to locate relevant facts in their consciousness. Notice how it significantly matters whether the agent is aware of alternative actions s/he might perform, and different possible outcomes or their rough likelihoods that different actions would bring about. (Sher 2009, 4–6.)

In what Sher considers an unrealistic, strong form, the searchlight view might be taken to mean that being fully responsible requires the moral agent to be aware of *all* relevant facts. This he takes to be quite absurd, as no one could then realistically be held responsible; given the cognitive and spatial limitations of human beings, and the relevant facts potentially including countless factors that we are all unaware of. Therefore, Sher's interest lies in a weaker, more realistic version of the searchlight view, where merely the amount of responsibility we can give to an agent is directly proportional to the range of relevant facts the agent was aware. (Sher 2009, 5.)

In what Sher considers the most plausible version of the searchlight view, the awareness of the agent does not require *active focus*, but merely *passive awareness*. This means that the agent need not be actively thinking about relevant aspects of his or her acts when performing them; it suffices that s/he is merely passively aware of them. Sher illustrates this by an example of driving a car: "A driver who is concentrating on not missing his exit may at the same time be aware that a car is approaching on his left, that he is hovering just over the speed limit, that his passenger is telling an anecdote, and of much else." Similarly, we are most often not actively focusing on, but merely passively aware of, aspects in our thinking that are relevant to our moral decisions. (Sher 2009, 6.)

Overall, what Sher considers the most plausible version of the searchlight view, and one that he is centrally arguing against, may be characterized as a view that holds that an agent's responsibility is directly proportional to their passive awareness of relevant facts of the situation.<sup>52</sup>

### 3.2 Indirect Evidence for the Appeal of the Searchlight View

Sher sees that the searchlight view has received little explicit defending anywhere but is still prevalent among academic philosophers in an implicit manner. He sees the view as being "the default position to which we gravitate when we are not thinking hard about the knowledge requirement" (Sher 2009, 7). This forces him to reconstruct the implicitly held view (that I presented in the previous section 3.1), and to build an indirect case to support his claim of the view being prevalent. To get some more perspective on the searchlight view, and on the sources where Sher sees the view being implicitly

held, it is helpful to examine where he sees it. He provides three pieces of indirect evidence to support the searchlight view being the default position among academic philosophers.

The first indirect piece of evidence comes from three citations by three preeminent philosophers representing three distinct traditions of philosophy: from Henry Sidgwick (a utilitarian), Jean-Paul Sartre (an existentialist), and Barbara Herman (a Kantian). Sher notes that each of these philosophers have assumed or presupposed the searchlight view; that an agent is responsible only insofar as s/he is aware of what s/he is doing. (Sher 2009, 7–8.)

The quotes by Sidgwick, Sartre, and Herman, that Sher presents (2009, 7–8), are, respectively:

“[The] proper immediate objects of moral approval or disapproval would seem to be always the results of a man’s volitions insofar as they were intended – i.e. represented in thought as certain or probable consequences of his volitions...” (Sidgwick 1981, 60.)

“The careless smoker who has through negligence caused the explosion of a powder keg has not *acted*. On the other hand the worker who is charged with dynamiting a quarry and who obeys the given orders has acted when he has produced the expected explosion: he knew what he was doing, or, if you prefer, he intentionally realized a conscious project.” (Sartre 1956, 433.)

“What an agent wills is a function of her grasp of a situation. If it is willings that are the object of moral assessment, judgments of right and wrong will then reflect the perspective of the agent, and so be relative to what she sees or considers relevant in the circumstances in which she acts.” (Herman 1993, 95.)

The second indirect piece of evidence that Sher notes is extracted from what he sees as common reactions to the phenomena of *negligence* and *culpable ignorance*. Sher gives a citation by philosopher Michael J. Zimmerman to illuminate this reaction: “all culpability can be traced to culpability that involves lack of ignorance, that is, that involves a belief on the agent’s part that he or she is doing something morally wrong” (Zimmerman 1997, 418).<sup>53</sup> This view Sher sees as misguided – which I shall address via Zimmerman in section 5.1. How this view indirectly supports the prevalence of the searchlight view is by it claiming that an agent’s control, and thus their responsibility, extends no further than their conscious choices. (Sher 2009, 8–9.)

The third and last piece of indirect evidence for the prevalence of the searchlight view is brought about via what Sher regards to be a common belief about an agent's responsibility. According to this common belief, "an agent is responsible only for those aspects of his behavior that are voluntary in the sense of being expressions of his *will*." Again, why this belief supports the prevalence of the searchlight view has to do with it boiling down to solely conscious phenomena (as opposed to also noting unconscious phenomena); namely the will appears to be essentially a conscious phenomenon. These concerns are more specifically outlined in section 3.4.2. (Sher 2009, 9.)

As a critical note, one may think – as, for example, Michael J. Zimmerman does (2009, 254) – that Sher's amount of evidence is quite thin to make the scale of generalization that he seems to be at times making; namely that the searchlight view has been *widely* accepted among academic philosophers. In fact, Zimmerman, who supports a *qualified* searchlight view, feels his views are in a minority among philosophers, let alone the general public (2009, 254). For what it's worth, Zimmerman may be right (see sect. 5.4). In any case, the importance of the topic seems to be already justified by merely noting that the searchlight view Sher reconstructs – or something relatively close to it – seems to have been implicitly accepted in at least *some* places often regarded in high value among academic philosophers. Still, for example, philosopher Angela Smith, who also opposes the searchlight view via what Sher calls *attributionism* (more recently also called *answerability*), has strongly agreed with Sher's characterization of the pervasiveness of the view (2010, 516; see also sect. 2.2). In any case, following Sher's reconstruction, "the searchlight view" has become part of the lexicon of moral philosophy, as, for example, Zimmerman seems to have adopted it to (roughly) describe his own position at least in some contexts, while distinguishing between two different kinds of searchlight views (Zimmerman 2009). I shall come back to Zimmerman's distinctions in section 5.1, and Angela Smith's views in section 5.2, where I present their evaluation of Sher's account.

### 3.3 Problems with the Searchlight View

Sher makes a case that if the searchlight view (i.e., his reconstruction) is to be considered correct, it would render everything outside of an agent's conscious control or awareness irrelevant to moral responsibility.<sup>54</sup> If the searchlight view is taken seriously, "then no one is responsible for any act, omission, or outcome whose moral or prudential defects can be traced to his lack of imagination, his lapses of attention, his poor judgment, or his lack of insight." In other words, an agent could not be held responsible without awareness. (Sher 2009, 7.)

In the following sections 3.3.1 and 3.3.2, I present Sher's basic objections to the searchlight view. The first objection has to do with how Sher sees the searchlight view conflating the engaged perspective and the detached perspective on action (3.3.1), and the second objection has to do with how our intuitions and practices concerning moral responsibility do not seem to match with the searchlight view (3.3.2). Sher seems to be primarily motivated via the intuition objection, but the objection concerning the engaged and the detached perspectives is helpful to introduce first, as his understanding of the two concepts seem central to his intuitions.

### 3.3.1 The engaged and the detached perspective

As Sher sees it, the fundamental problem with the searchlight view, and the main reason why it should be rejected – and why a better view should be formulated – has to do with it conflating two different and incompatible perspectives on action (Sher 2009, 9):

- 1) The engaged perspective (a first-person perspective): the perspective we occupy when we ourselves act.
- 2) The detached perspective (a third-person perspective): the perspective we occupy when we consider people's acts – our own or those of others – “from the outside.”

Sher summarizes:

“My reason for thinking that the searchlight view conflates these perspectives is that it draws on a conception of the agent that derives its plausibility from the first perspective while purporting to specify a necessary condition for the applicability of a concept – responsibility – whose natural home is the second.” (Sher 2009, 9.)

*The engaged perspective* – or what Sher also calls *the deliberative perspective* – is the usual perspective we take when engaging the world as agents. Every time we make decisions, consider reasons for and against performing different acts, and consider different outcomes, the process is always fully conscious – that is, on the level of conscious experience. We may be aware that we have many beliefs, predispositions, and other psychologically influential characteristics that we are unaware of, but on the level of conscious experience these are inaccessible to us, and hence irrelevant to the practicalities of engaging the world. We thus perceive ourselves, from the engaged perspective, as narrowly bounded by our own consciousness; we experience ourselves as (volitionally effective)

center of consciousness. This easily leads us to think, rather straightforwardly, that we cannot be responsible for something that is not consciously experienced. (Sher 2009, 10, 41–42.)

*The detached perspective*, on the other hand, is the perspective from which responsibility towards others usually arises. The judgment of whether someone is responsible for a particular act presupposes that the act in question has already been performed. We can engage in this perspective also concerning ourselves, in retrospect – a very different perspective than the one we had when we were still either deliberating about the act to be done or when we were performing the act. From this perspective, the practical question does not concern whether we should or shouldn't perform a particular act, but at most how we should react to its performance. In the detached perspective, we must also consider all potentially relevant facts that we have managed to access after the act. Unlike from the engaged perspective, from the detached perspective “there is no *a priori* reason to exclude those states of the agent that were not illuminated by the searchlight of his consciousness.” (Sher 2009, 10.)

The engaged perspective and the detached perspective are incompatible, according to Sher, in the sense that no one can simultaneously occupy them both. We are always operating under one or the other, depending on whether the act is being deliberated or having already been done. Still, Sher emphasizes that we could still occupy one perspective while taking account of how things appear from the other. This means that even though our detached perspective always examines responsibility from an external point of view, our assessments via that view could still greatly – even fully – depend on an agent's internal and engaged perspective; and, thus, of their conscious experience. In other words: we may still want to and have good reasons to take note of the engaged perspective via occupying the detached perspective. This sparks the question that Sher regards to be one on which the tenability of the searchlight view depends on: *Are there any considerations that compel us to ignore all cognitive states to which the agent lacked access to, even if our external perspective does not itself do so?* (Sher 2009, 10–11.)

Some relevant considerations include:

“...the practical nature of the concept of responsibility, the unfairness of blaming or punishing people for what they cannot help, the connection between imaginatively identifying with an agent and understanding what he has done, and the unreasonableness of any demand with which the agent to whom it is directed is unable to comply.” (Sher 2009, 11.)

Sher argues that none of these considerations ends up supporting the notion that holding an agent responsible requires regarding what s/he did exclusively from their own perspective (I examine these arguments more closely in section 3.4).

Overall, Sher's intention seems to be to demonstrate that as the searchlight view seems to consider *only* the conscious aspects of an agent to be relevant for moral responsibility, it cannot be an adequate view. At the very least it is his intention to undermine the searchlight view's appeal.

### 3.3.2 Intuitions and practices

A substantial and apparently motivating premise behind Sher's case for the searchlight view being inadequate is that it sometimes goes against what he considers to be our common intuitions about responsibility. Relatedly, it doesn't match with what he considers our actual practices concerning responsibility. He sees that we tend to hold people – others and ourselves – responsible for *lack of imagination, lapses of attention, poor judgment, and lack of insight* in various situations. He sees that we tend to hold agents responsible from the detached perspective, for cognitive states the agents lacked access to – as opposed to just exclusively noting their past engaged perspective, and hence only the cognitive states that they did have access to (as the searchlight view seems to do). Thus, ultimately, Sher argues that as our intuitions and practices concerning responsibility sometimes conflict with the searchlight view, it is in some way lacking (specifically by means of conflating the engaged and the detached perspective, as introduced in the previous section 3.3.1). (Sher 2009, 9–12; Waller 2014, 641.)

Sher backs up his premise that our common intuitions go against the searchlight view, in some cases, via nine example cases.<sup>55</sup> The cases are designed to illustrate that in some situations our intuitions would say that an agent is responsible without awareness, and that thus the searchlight view is inadequate. The example cases aim to illustrate this via three ways in which the searchlight view seems to fail: according to Sher, it fails to capture our intuitions in situations where an agent acts wrongly because s/he (1) loses track of some crucial elements of their situation, (2) exercises poor judgment, or (3) lacks in moral insight or imagination (Sher 2009, 23).<sup>56</sup> More colloquially, Sher briefly summarizes these three to concern (1) forgetting, (2) bad judgment, and (3) insensitivity (Sher 2009, 31). In his review of Sher's book, Michael J. Zimmerman further summarizes these three aspects as roughly having to do with (1) involuntary lapses of judgment, (2) poor judgment, and (3) lack of moral insight (Zimmerman 2009, 249).

The first three of the nine Sher's example cases correspond to the first kind of situations where our intuitions, according to Sher, fail to match with the searchlight view: situations that involve an *involuntary lapse of judgment*. The cases Sher presents are (quoted directly from Sher 2009, 24):

1. *Hot Dog*. Alessandra, a soccer mom, has gone to pick up her children at their elementary school. As usual, Alessandra is accompanied by the family's border collie, Bathsheba, who rides in the back of the van. Although it is very hot, the pick-up has never taken long, so Alessandra leaves Sheba in the van while she goes to gather her children. This time, however, Alessandra is greeted by a tangled tale of misbehavior, ill-considered punishment, and administrative bungling which requires several hours of indignant sorting out. During that time, Sheba languishes, forgotten, in the locked car. When Alessandra and her children finally make it to the parking lot, they find Sheba unconscious from heat prostration.
2. *On the Rocks*. Julian, a ferry pilot, is nearing the end of a forty-minute trip that he has made hundreds of times before. The only challenge in this segment of the trip is to avoid some submerged rocks that jut out irregularly from the mainland. However, just because the trip is so routine, Julian's thoughts have wandered to the previous evening's pleasant romantic encounter. Too late, he realizes that he no longer has time to maneuver the ferry.
3. *Caught off Guard*. Wren is on guard duty in a combat zone. There is real danger, but the night is quiet. Lulled by the sound of the wind in the leaves, Wren has twice caught herself dozing and shaken herself awake. The third time she does not catch herself. She falls into a deep slumber, leaving the compound unguarded.

The second three cases are ones in which the agents, according to Sher, seem responsible without awareness for wrong acts they perform because they display *poor judgment*, thus corresponding with the second kind of situations where he sees our intuitions to fail to match with the searchlight view. These are as follows (quoted directly from Sher 2009, 26):

4. *Home for the Holidays*. Joliet, who is afraid of burglars, is alone in the house. Panicked by sounds of movement in her kitchen, she grabs her husband's gun, tiptoes down the stairs, and shoots the intruder. It is her son, who has come home early for the holidays.
5. *Colicky Baby*. Scout, a young woman of twenty-three, has been left in charge of her sister's baby. The infant is experiencing digestive pains and has cried steadily for hours. Scout has made various attempts to ease its discomfort, but nothing has worked. Finally,

to make the child sleep, she mixes vodka with its fruit juice. The child is rushed to the hospital with alcohol poisoning.

6. *Jackknife*. Father Poteet, a good driver, is gathering speed to enter a busy freeway. Because the merge lane is very short, he must either pull in front of a looming eighteen-wheeler or stop abruptly. He makes the split-second decision that he has room to merge, but he is wrong. The trucker hits the brakes hard, his truck jackknifes across four lanes of traffic, and many people are seriously injured.

The last three of the nine Sher's example cases are ones which involve *lack of moral insight or imagination*, thus corresponding with the third kind of situations where, according to him, our intuitions seem to fail to match with the searchlight view. These are the following (quoted directly from Sher 2009, 28):

7. *Bad Joke*. Ryland is very self-absorbed. Though not malicious, she is oblivious to the impact that her behavior will have on others. Consequently, she is bewildered and a bit hurt when her rambling anecdote about childless couple, handicapped person, and a financial failure is not well received by an audience that includes a childless couple, a handicapped person, and a financial failure.
8. *Bad Policy*. Sylvain, a college professor, is empathetic to a fault. He identifies with troubled students and freely grants their requests for opportunities to earn extra credit. Because he enters so completely into each interlocutor's perspective, he often forgets that there are other less aggressive students who would eagerly welcome the same chance. As a result, his grading policy is inconsistent and unfair.
9. *Bad Weather*. It is 1968, and amerika (a nom de guerre) is a member of the Weather Underground. Sensitive and conscientious as a child, amerika has been rethinking his moral beliefs. In a series of stages, he has become convinced, first, that capitalism is deeply unjust; next, that nothing short of revolution will bring change; and, finally, that the need to rectify massive injustice far outweighs the rights or interests of mere individuals. To procure funds for the Revolution, amerika takes part in a robbery in which a bank guard is killed.

Sher further summarizes the particulars in the nine example cases:

“We have now encountered a total of nine cases in which agents seem responsible for wrong acts whose wrongness they did not recognize. Although the agent's lack of awareness is crucial

to all nine cases, it does not always take the same form. In three cases (*Hot Dog*, *On the Rocks*, *Caught off Guard*), the agent does not even realize that he is in a situation that calls for action, while in the remaining six he does realize this but lacks an accurate appreciation of what he ought to do. In one of these six cases (*Home for the Holidays*), the agent's cognitive defect is due to a distorting emotion (panic), in two others (*Bad Joke*, *Bad Policy*), it can be traced to his insensitivity to a morally relevant factor, and in the remaining three (*Colicky Baby*, *Jackknife*, *Bad Weather*) it is a product of unadorned poor judgment. In two of the latter cases (*Colicky Baby*, *Jackknife*), the defective judging occurs when the agent is assessing the facts, while in the third (*Bad Weather*), it occurs when he is thinking through his moral beliefs. Because the patterns of error are quite diverse, and because each one is instantly recognizable, the range of counterexamples to the searchlight view can already be seen to be broad." (Sher 2009, 28–29.)

Thus, overall, according to Sher, the searchlight view is an inadequate description of our intuitions and practices concerning moral responsibility. In many cases the searchlight view suffices to account for our intuitions, but apparently there seems to be situations where the searchlight view fails to account for them. Namely, as illustrated, these are situations where we seem to deem agents morally responsible without awareness for their (1) involuntary lapses of judgment, (2) poor judgment, and (3) lack of moral insight (Sher 2009, 23; Zimmerman 2009, 249). According to Sher, in most of the example cases "the agent would definitely be blamed and might well be liable to punishment" (Sher 2009, 24–28).<sup>57</sup>

Of course, the possible existence of these kinds of cases doesn't settle the question of whether the searchlight view should be expanded upon. Rather, the question now becomes: should we try to adjust our intuitions to better match with the searchlight view or something similar, or should we try to accommodate these kinds of intuitions within our thesis concerning the role of awareness, and knowledge, in moral responsibility? (Sher 2009, 24, 33–39; Zimmerman 2009, 250.)

### 3.4 Imaginative Reconstruction of Failures to Justify the Searchlight View

The question of whether to question our seemingly common intuitions or to question the searchlight view is decided by examining which of the two options can be better justified via a stronger argument. Sher advocates the latter. To justify his case, he critically examines attempts that could be interpreted as trying to justify something like the searchlight view and why he considers those attempts to not

succeed. His examination relies practically fully on his own “*imaginative reconstruction*”, because – as seen in section 3.2 – he sees that there has been strikingly little direct attention given to the often implicitly held searchlight view. (Sher 2009, 41–42.)

His working hypothesis is the assumption that the best defense of the searchlight view is likely to involve some premise that emphasizes the engaged – i.e., deliberative – perspective. Thus, he examines the question proposed earlier in section 3.3.1: *Are there any considerations that compels us to ignore all cognitive states to which the agent lacked access to, even if our external perspective does not itself do so?* If there can be found such considerations, it would give more support for us critically examining our intuitions, but if there cannot be found such considerations, it in turn would give more support for us critically examining and expanding on the searchlight view. (ibid.)

As stated earlier, in section 3.3.1, relevant considerations include (1) the practical nature of the concept of responsibility, (2) the unfairness of blaming or punishing people for what they cannot help, (3) the connection between imaginatively identifying with an agent and understanding what s/he has done, and (4) the unreasonableness of any demand with which the agent to whom it is directed is unable to comply (Sher 2009, 11). In this section, I present Sher’s arguments for why none of these considerations ends up supporting the notion that holding an agent responsible requires regarding what s/he did exclusively from their own perspective (at least insofar as Sher himself has imagination to fairly assess these sorts of considerations, based on his own reconstructions that seem to lack explicit arguments by others). Sher connects the consideration 1 to considerations about responsibility as a concept of practical reason, which I outline in section 3.4.1. Considerations 2, 3, and 4 he connects to Kantian considerations about fairness, outlined in section 3.4.2.

At face value, both of the following lines of consideration appeal to the fact that we can only deliberate about those features of the available actions of which we are aware; and thus, potentially implicitly support the notion that responsibility is bounded only to the conscious aspects of an agent – and thus, potentially implicitly support the searchlight view (Sher 2009, 17). Here, following Sher, I move on to refer to the engaged perspective solely as the *deliberative perspective*, to emphasize the fact that it appears to be the perspective in which we can deliberate about our antecedent reasons for acting.

### 3.4.1 Responsibility and practical reason

As the first candidate that might be considered to make a strong implicit case for the searchlight view, Sher examines the idea that responsibility is a practical concept – i.e., falling under *practical reason*

(Sher 2009, 41). “Practical reason” refers to a practical standpoint of reflection, characterized by normative questions; with matters of value, like what one ought to do, or what it would be best to do. In this standpoint, emphasis is put on evaluating *reasons for action* (or omission) from first-person point-of-view and directed towards future acts (reminiscent of the deliberative perspective). The contrasting standpoint would be “theoretical reason”, characterized by us engaging in reasoning that is directed at the resolution of questions that are in some sense theoretical rather than practical; for example, reasoning about questions of explanation and prediction. In the theoretical standpoint, emphasis is put on *reasons for belief*, matters of fact and their explanation, and from a more so third-person perspective (reminiscent of the detached perspective).<sup>58</sup> (Wallace 2018, ch. 1.)

Sher explains that the reason why some consider responsibility a practical concept is that it is precisely when we deliberate that we are compelled to regard ourselves as responsible for whatever choices we make. And as responsibility would be a practical concept, the preconditions for responsibility would also derive from the deliberative perspective, and thus, potentially, appealing to the notion of the searchlight view that responsibility is bounded only to the conscious aspects of an agent. (Sher 2009, 17–18.)

To be clear, Sher himself thinks the claim that the concept of responsibility would be practical is deeply problematic. The problem he sees in the notion is that if responsibility is a practical concept, then it would be logically impossible to deliberate about anything except one’s own future actions, which would go against what he considers our “ordinary concept of responsibility” where responsibility seems to be neither oriented to the future nor restricted to the first-person point-of-view. However, as we see below, this doesn’t rule out the *possibility* that practical deliberation about our future actions could lead to putting emphasis also to our past actions and to actions of others, in terms of application of responsibility. (Sher 2009, 45–46.)

To examine the account of responsibility falling under practical reason, and whether support for the searchlight view might be found there, Sher examines the accounts of philosophers Christine Korsgaard and Hilary Bok, both of whom advocate responsibility being a practical concept. Korsgaard’s and Bok’s arguments are different, albeit not in a way that would make them mutually exclusive. Sher examines both arguments separately.<sup>59</sup> (Sher 2009, 41–54.)

According to Sher’s account, Korsgaard and Bok both exploit the Kantian insight that all deliberation takes place “under the aspect of freedom” to connect the deliberative perspective to the concept of responsibility (albeit via different arguments). Even if we are fully determined, the

Enlightenment era philosopher Immanuel Kant (1964, 116) views that we are incapable of thinking ourselves as incapable of making a difference when we deliberate. This line of theory, potentially, could support the searchlight view, if the deliberative perspective is connected to responsibility in such a way that it would be the only thing to consider, i.e. that it would disregard the detached perspective. Relevantly, advocating responsibility as practical reason, both Korsgaard and Bok have denied that the point of regarding someone as responsible is either to describe him/her or to explain what s/he has done. Thus, the theory can be seen to potentially subscribe to the searchlight view.

However, Sher argues that Korsgaard would likely not accept the searchlight view, while Bok's views on the matter remain dependent on unanswered questions. He arrives to this conclusion by examining the ways in which Korsgaard and Bok have attempted to show that a practical concept of responsibility is not only applicable to our own future acts, but also to (a) acts that we have already performed and similarly to (b) the acts of others. These are what Sher refers to as the *future-past gap* and the *self-other gap* in the theory of responsibility as practical reason, respectively. For practical responsibility to not conform to the searchlight view, these gaps would need to be bridged, as bridging them would enable practical concept of responsibility to apply also to our past and others' past acts (i.e., to apply to the detached perspective, not only the deliberative perspective) – also potentially in a way that would allow responsibility without awareness. (Sher 2009, 42–45.)

Regarding the future-past gap: According to Sher, Korsgaard considers the mutual or reciprocal quality of practical reason crucially important, because genuine mutuality can obtain only among temporally enduring agents. Following this, Sher reads Korsgaard to implicitly argue – via her larger interpersonal theory of practical reason – that responsibility as a practical concept can be applied to our past acts, because for us to be reciprocal agents, or the kind of agents with whom others can act in concert, we must be willing to take responsibility for our past actions no less than for our future ones. In other words: it seems that our future actions in terms of reciprocal success compel us to take responsibility for our past acts, thus bridging the future-past gap. (Sher 2009, 46–47.)

Bok's argument, on the other hand – as described by Sher – is that we are responsible for past acts not (necessarily) because it would relate to our reciprocal ability, but because it relates to the adequacy of our own future decisions. Bok seems to say that our past acts reveal features of our character that *influence* our future decisions, and thus responsibility as a practical concept would also compel us to note our past actions so we can better evaluate our future actions and, if seen fit, find reasons to change our relevant features and behavior. (Sher 2009, 47–48.)

After finding out that Korsgaard and Bok have essentially provided initial answers to the future-past gap, Sher proposes some critical questions to their accounts of bridging the gap.<sup>60</sup> However, Sher concludes that a further treatment is not needed in this context, because even if Korsgaard's and Bok's accounts successfully show that a practical concept of responsibility applies to both past and future acts, the premises of that conclusion do not support an omnitemporal version of the searchlight view. In other words: it suffices that the future-past gap seems to have been bridged in the philosophers' premises, even if there are further questions about the nuances of their conclusions or lines of argument. As Korsgaard and Bok argue that practical reason compels us to take responsibility for our actions after we perform them – due to reciprocity (Korsgaard) or due to influencing our future decisions (Bok) – then our responsibility seems not to be necessarily limited to what we were aware of but may also include what we were *unaware* of. Even if responsibility is regarded as a practical concept, at least Korsgaard's and Bok's accounts do not seem to presuppose that responsibility is limited to features we were only antecedently aware of. (Sher 2009, 48–50.)

But what about the self-other gap? There, Sher finds a similar conclusion. He interprets Korsgaard to argue that a practical concept of responsibility applies not only to our own acts but also to acts of others, because that enables reciprocal relations with others. As it was with the future-past gap, the answer here also relates to Korsgaard's larger interpersonal theory of practical reason. Reciprocity, central to Korsgaard's understanding of practical reason, requires both that we view other people's ends as reasons because they seek them and that we expect the others to view our own ends as reasons because *we* seek them. Thus, we have a reason to regard other people just as capable of evaluating their reasons, and hence regard them just as responsible for any decisions we would consider ourselves to be. And thus, a practical concept of responsibility is not only transtemporal (bridging the future-past gap) but also intersubjective (bridging the self-other gap). Furthermore, to emphasize Korsgaard's willingness to reject the searchlight view, Sher notes that she has herself explicitly stated that we “may well blame agents for involuntary attitudes or expressions, because we blame people for lack of control itself” and that we ought to do so (Korsgaard 1996a, 198). (Sher 2009, 51–53.)

As Sher presents it, Bok's argument for a practical concept of responsibility bridging the self-other gap is that we have good reason to regard others as just like us, as “persons who are capable of governing their lives through practical reasoning” (Bok 1998, 188). She states that “[w]hen we regard others as persons, we must both extend to them a view we would otherwise take of ourselves and allow ourselves to enter into their view of themselves” (Bok 1998, 188). These premises, together with Bok's earlier conclusion that practical reason compels us to view ourselves responsible (and the

further premise that each other person's practical reason can be seen to impose the same demands on them), Sher describes to lead Bok to conclude that our practical concept of responsibility extends to others as well as ourselves. (Sher 2009, 53.)

Similar to Korsgaard and Bok bridging the future-past gap, Sher also has critical questions concerning their bridging of the self-other gap. But, once again, Sher concludes that a further treatment is not necessary in this context, for the philosophers' premises already do not seem to presuppose the searchlight view. Even a practical concept of responsibility seems not to be necessarily limited to neither what we are aware of, nor to what we were aware of, nor to what others are or were aware of, but may also include what we were *unaware* of and what others were *unaware* of. Practical concept of responsibility is thus not necessarily limited to features we, or others, were only antecedently aware of. In other words: practical concept of responsibility seems not to be confined to only the deliberative perspective but may also utilize the detached perspective – and hence does not presuppose the searchlight view. (Sher 2009, 52–54.)

In summary, Sher's conclusions on his examination of responsibility as a practical concept is: (1) the viability of Korsgaard's and Bok's attempts to bridge the future-past and self-other gaps depends on the answers to various questions that remain unanswered, but that (2) even if those attempts succeed, they do not end up supporting the original rationale that a practical concept of responsibility applies only to features of acts of which agents are aware. Thus, theorists who count responsibility under practical reason seem not to be – at least *a priori* – committed to the view that solely the requirements of the deliberative perspective should be noted in responsibility, and thus they seem not to necessarily subscribe to the searchlight view. (Sher 2009, 18; see also Zimmerman 2009, 250–251.)

### 3.4.2 Responsibility and the Kantian Principle

After having argued that the searchlight view does not follow from the practical reason theorists' premise that the concept of responsibility originates from the deliberative perspective, Sher notes that an alternative route to implicitly support the searchlight view still exists. Namely, the principle that holding agents responsible for what they did not foresee is *unfair* because their lack of foresight deprived them of *control*. Sher describes this principle as often thought to be the basis for Immanuel Kant's claim that an agent's moral value depends entirely on the quality of the agent's will – that a good will, even if ineffective, would “shine like a jewel for its own sake” (Kant 1964, 62).<sup>61</sup> This

connection to Kant prompts Sher to refer to the principle as *The Kantian Principle*. (Sher 2009, 55–56.)

Sher takes the Kantian Principle to be very widely accepted. To demonstrate this, he quotes three contemporary philosophers who have referred to the principle in the contexts of discussing responsibility, free will, and moral luck, respectively (Sher 2009, 56):

“It is often said to be unjust to blame someone for what he could not help doing.... We think it unfair to adopt an attitude of disapproval toward someone on account of an act or omission, where this was something outside his control.” (Glover 1970, 70 & 73.)

“It is an illicit generalization to infer that having alternative possibilities is never required for moral responsibility or free will – throughout an entire lifetime. One could claim this only if one could claim that it is fair to hold persons responsible for being what they are even though there is nothing they could ever have done to make themselves different than they are.” (Kane 2002, 697.)

“If morality depends on luck, then at least sometimes people are judged morally for things that are beyond their control. This seems to be unfair; one does not deserve to be held responsible for what is beyond one’s control.” (Statman 1993, 2–3.)

Even though the Kantian Principle clearly seems to enjoy wide acceptance, Sher states that the principle is most often either merely presupposed than explicitly stated, or baldly asserted than carefully defended. For example, none of the quoted philosophers, according to Sher, even hint at a defense where the principle’s notions of “control” and “fairness” would be explicated.<sup>62</sup> This, for Sher, leaves a lot of room for interpretation in terms of whether the principle could fit in with the searchlight view or reject it. Particularly, as seen before, Sher is interested whether, and if so why, the only relevant facts about an agent are those concerning their conscious choices.<sup>63</sup> He asks: “[w]hy shouldn’t they also include facts about the relation between [the agent’s] acts and his unchosen vices... his native talents or limitations... and facts that he has learned but cannot at the moment retrieve?” As no explicit let alone settled account for a conception of “control” is found, Sher considers there to be a lot of room for a conception that doesn’t take it to require conscious choice. In fact, there may be room for many equally defensible conceptions of control, but for Sher it would suffice that the proponents of the searchlight view could demonstrate that those versions of the Kantian Principle that take conscious awareness to be a necessary condition for control can defend

such a view. Or, alternatively, it would suffice that the opponents of the searchlight view demonstrate those versions to be *indefensible*. (Sher 2009, 56–58.)

In the following sections 3.4.2.1–3.4.2.3, I go through Sher’s imaginative reconstruction of the Kantian Principle step-by-step – which he conducts in pursuit of a version that would take conscious awareness to be a necessary condition for control, and that would thus support the searchlight view. In section 3.4.2.4, I try to very briefly summarize the complex reconstruction along with Sher’s conclusions. The next three sections illustrate the complexity of Sher’s reconstruction; but for a busy reader, the last section, 3.4.2.4, likely suffices to provide the basic outline of Sher’s reconstruction and conclusions.

#### 3.4.2.1 *The Kantian Principle and imaginatively identifying with an agent*

Engaging in his imaginative reconstruction, Sher outlines that a version of the Kantian Principle that would take conscious awareness to be a necessary condition for control – and to which proponents of the searchlight view could thus appeal to – would return to demands of the deliberative perspective. From the deliberative perspective, we can only appeal to facts of which we are aware, and thus a deliberating agent’s conscious beliefs – as opposed to unconscious elements of their psychology – must be central both to their conception of what is within their *control* and to their conception of what they may *fairly* be asked to do. Thus, from the perspective of deliberating agents, the only form of control that matters, when judging fairness of someone blaming them or holding them responsible, is *searchlight control*: the ability to locate relevant facts in their consciousness. (Sher 2009, 58–60.)

To clarify, Sher summarizes the features of the resulting version of the Kantian Principle, to which he initially thinks proponents of the searchlight view could appeal to:

“[T]he form of control in whose absence it is unfair to hold an agent responsible will simply be searchlight control, while the sense in which it is unfair to hold an agent responsible for a feature of an act over which he lacked searchlight control will simply be that in which he himself would have regarded as unfair a demand that he perform or not perform an act with that feature. ... Moreover, because we understand what is wrong with being unreasonable, we also seem to understand this version’s normative force.” (Sher 2009, 60.)

However, in terms of the proposed version of the Kantian Principle truly being one that the proponents of the searchlight view could appeal to, Sher sees a complication. Namely: It is one thing for an agent

to *regard* something as unfair, and another thing for the Kantian Principle to *condemn* something as unfair. What an agent is compelled by their deliberative perspective to regard as unfair are certain *demands* directed at them, which they cannot see how to fulfil – whereas, by contrast, what the Kantian Principle condemns as unfair are certain ways of reacting, e.g. *blaming and attributing responsibility*, to an agent when the agent fails to conform to demands that they cannot see how to fulfil. In Sher's account, the searchlight view would seem to want to justify the latter sense of unfairness with the former. But the unfairness of blame or attribution of responsibility (an act of the detached perspective) cannot be automatically inferred from the unfairness of the demand itself (an experience relating to the deliberative perspective). Furthermore: for Sher, it seems that there may be *two concepts of "unfair"* at play here, because blaming and holding people responsible are not themselves demands. Thus, the following question becomes important: "Exactly why *should* holding someone responsible for failing to fulfil a demand that he was bound to view as unfair be viewed as unfair in any further sense?" (Sher 2009, 60–61.)

For Sher, this puts the proposed version of the Kantian Principle into serious doubt: Even if the people we blame or hold responsible are unable to appeal to any considerations beyond their conscious beliefs when they made their decision, and would regard our ways of reacting by blame or attribution of responsibility as unfair, there isn't any obvious reason for us to restrict *our* attention when considering their decision after their act. As, according to Sher, these are two different perspectives, there isn't any obvious reason to *not* base our blame or attributions of responsibility on features of the agent and their past choice situation that they were not aware. (Sher 2009, 61.)

This doubt could only be resolved if it were found that blaming and holding people responsible actually *are* reactions that happen from the perspective of the target agent. To rule out this possibility, to his own satisfaction, Sher examines philosopher Thomas Nagel's suggestion that this would be the case. Nagel's view, as Sher seems to accurately describe it, is that blaming or holding someone responsible necessarily involves imaginatively reconstructing both the agent's choice situation from their perspective and the reasons that the agent used to arrive to their choice (Nagel 1986, 120). If Nagel is right, this process of *imaginatively identifying with an agent and understanding what s/he has done* could compel us to share the agent's view that any demands that they could not see how to meet are unfair because unreasonable. (ibid.)

Sher conjectures that Nagel doesn't mean we ought to *fully* identify with an agent, merely take account of their perspective. However, Sher reasons that a merely partial-identification view would not resolve the gap between the two perspectives of the agent and the accuser. If merely a partial identification with the agent is required, this would not compel us, as the accuser, to share the agent's

view in a way that would nullify us basing our blame or attribution of responsibility on features of the agent and their past choice situation that they were not aware. And if a full-identification view was proposed, then no one could ever blame or hold any (non-akratic) agent responsible, because full identification with the agent would always require us to view that the agent had the most reason to do exactly what they *did*. Thus, Sher reasons that blame and attribution of responsibility require a degree of detachment from the persons at whom they are directed, and thus this line of reasoning doesn't support the proposed understanding of the Kantian Principle being one that could actually support us taking note only of the deliberative perspective. And hence, it fails to provide an understanding of the principle that the proponents of the searchlight view could appeal to. (Sher 2009, 61–63.)

#### 3.4.2.2 *Fair and unfair reactions, and retrospective endorsement of a demand* – another way to approach “unfairness” in the Kantian Principle

And so, Sher's challenging imaginative reconstruction continues, as an answer to the dilemma was not found. The dilemma being: “if the agents themselves must regard as unreasonable, and thus as unfair, all demands that they could not see how to fulfil, in what sense, and why, is it also unfair for us to blame or hold them responsible for not *fulfilling* such demands?” (Sher 2009, 63.)

A further distinction Sher makes is between two types of (detached) reactions to an agent's failure to fulfil a demand that they would regard as an unfair demand: reactions that are themselves *not* unfair, and reactions that are. Some examples of reactions that do not seem unfair, that Sher mentions, include feeling sorry for the agent, regretting the limits of what the agent realized about their situation, or trying to enlighten the agent about what they did not realize. Even if the agent would regard the *a posteriori* demand unfair, these reactions seem not unfair themselves. But the proponents of the searchlight view might view that some other reactions, namely blaming and holding the agent responsible, do seem unfair. Sher reasons that what distinguishes the reactions that are not themselves unfair from the unfair reactions is the latter committing us to *endorse* the moral demand that we take the agent to have failed to satisfy. The former examples do not endorse any demand that the agent would have regarded as unfair, whereas blaming or holding someone responsible implies that it *was* legitimate to demand that the agent do something other than what they did. (Sher 2009, 63–64.)

Herein, Sher proposes a second way to understand the Kantian Principle, in terms of it being a version to which the proponents of the searchlight view could appeal to: As blame and attribution of responsibility involve the retrospective *endorsement* of the demand the agent is believed to have

failed to meet, and as endorsing a demand is in many ways similar to making a demand, the conception of “unfairness” that applies to the demand could also apply to its retrospective endorsement. Thus, the newly understood concept of “unfair” in the Kantian Principle, would focus on the retrospective-endorsement component. This, Sher proposes, may bridge the gap between the sense of “unfair” that refer to demands, and the sense that refer to blame and attribution of responsibility. (Sher 2009, 64.)

Having now formulated an understanding of the Kantian Principle that, to his tentative satisfaction, the proponents of the searchlight view could potentially appeal to, Sher remains vigilant about the distinction between the detached and the deliberative perspective. This causes him to hesitate to accept the resulting version of the Kantian Principle. He asks:

“Why, exactly, should a conception of unfairness that has its natural home in the deliberative perspective have a normative claim on anyone who does not occupy that perspective? It is clear enough why someone who is told to do something that he does not see how to do must *himself* regard that demand as unfair, but nothing yet said explains why some other person, who does not share his perspective, must take a similar view. This last question – Why should the perspective of the agent be as authoritative for others as it is for him? – remains the central challenge to our ability to defend an interpretation of the Kantian Principle that supports the searchlight view.” (Sher 2009, 64–65.)

As illustrated earlier, in section 3.3.1, Sher regards the natural home of responsibility to be the detached perspective, which seems to conflict with the conclusion that the natural home of unfairness would be in the deliberative perspective. Nevertheless, Sher sees there to be a possible way to meet the abovementioned challenge he himself raised. Namely, by noting that demands are by their nature constrained by the perspectives of those whom they address, and thus the demand must be accessible enough to the agent, if we want to induce relevant changes in their practical deliberation. Violations of this requirement are characterized by not taking note of the agent lacking some crucial information about what we would like to demand of them. As an example, Sher mentions how a parent is unable to respond to a demand that he stops endangering his child if he is unaware that, or of how, he is endangering his child. In such a situation, the agent would regard the demand as unfair because unreasonable. Thus, a conclusion that the agent’s perspective is authoritative in determining which demands are unfair may seem reasonable, and that anyone who blames or holds an agent responsible should endorse only demands accessible to the agent. (Sher 2009, 65–66.)

### 3.4.2.3 “Accessibility” and the aim of moral demands

Despite having established a more precise – though still tentative – understanding of a version of the Kantian Principle that the proponents of the searchlight view might appeal to, Sher remains skeptical this succeeds in establishing a version that would support the searchlight view. The problem he sees is in its premise that ‘no demand can influence an agent’s deliberation unless it is accessible to them’. (Sher 2009, 66.)

For the Kantian Principle to support the searchlight view, Sher underlines that it must be understood to assert that it is unfair to blame or hold an agent responsible for anything they did not consciously choose. For the now proposed version of the principle to accomplish this, it would need to “equate access with conscious awareness... [and] assert that no demand can influence any agent’s deliberation unless the agent has access to the demand *in the sense of consciously believing that it calls for a specific act that he sees himself as able to perform.*” But, interpreted this way, Sher sees the premise to be false, because agents, in their practical deliberation, can be and are constantly being influenced by normative demands that they are not consciously aware of. For some examples, Sher mentions: when an agent automatically does their duty; whenever an agent does not even consider lying in a situation where it would benefit them; or whenever an agent unthinkingly avoids a hurtful topic in a conversation. (Sher 2009, 66–67.)

However, Sher’s further solution is to not interpret accessibility in terms of conscious awareness, but rather in terms of the demand’s ability to make effective contact with the agent’s whole complex cognitive and motivational system. But in this sense, Sher reasons, the premise becomes too weak to support the requisite version of the Kantian Principle. This is because this interpretation does not imply that every demand of which the target agent is unaware fails to make an impact, i.e. fails to achieve its *internal aim*. Thus, it does not imply that all demands of which the target agent is unaware are unreasonable – or that the retrospective endorsement of any demand is unreasonable. (Sher 2009, 67.)

Sher then continues to explore an option that the demands could be understood as unreasonable if the internal aim of demands is understood to be something else than to simply elicit the demanded action (that he has assumed thus far). Alternatively, the internal aim of demands – at least demands that blame and attribution of responsibility seem to imply – could be understood to not only elicit the demanded action but to elicit it *for the right reasons* (Sher 2009, 67–68). But even under this assumption, Sher notes:

“...the mere fact that someone does not consciously register the reason-giving aspects of his situation does not mean that the corresponding moral demand cannot fully achieve its aim... [Thus,] retrospectively endorsing a demand of whose factual basis the target agent was not aware – and so, by extension, blaming or holding someone responsible for not meeting that demand – is not necessarily unreasonable.” (Sher 2009, 68.)

One final possibility that Sher explores, that might turn the Kantian Principle to support the searchlight view, is that the internal aim of moral demands was understood yet more stringently:

“[It could be understood that the demands] aim not merely at inducing agents to perform for the right actions, nor yet at inducing agents to perform the right acts for the right reasons, but rather at inducing agents to perform the right actions for the right reasons *through the recognition of those reasons*.” (ibid.)

Sher views that philosopher Thomas Scanlon – who endorses attributionism, which will be explored more in sections 4.1.3, 5.1, and 5.2 – comes close to endorsing this sort of proposal, even though Scanlon’s own topic is not exactly the internal aim of moral demands. Scanlon’s view seems to be that conscious recognition of features of one’s situation that provide reasons to do something is required for being conscious of the reasons to do something (Scanlon 1998, 290). For example, to turn to Sher’s earlier example: for a parent to be aware of the reasons to stop endangering his child, he needs to consciously recognize the relevant features of his situation (i.e., he needs to recognize the features that endanger his child, and the reasons to stop endangering his child). (Sher 2009, 68–69.)

This, Sher takes as the most promising line of understanding the Kantian Principle, in terms of it supporting the searchlight view, but doesn’t find any persuasive reasons why we should accept such a stringent view of the internal aims of moral demands. Specifically, Sher doesn’t find two possible reasons persuasive, both of which could be described Scanlonian:

“The first, which takes as its point of departure Scanlon’s central idea that an act’s rightness depends on its conformity to principles that no one could reasonably reject, is that any attempt either to arrive at such a set of principles or to justify one’s conduct in terms of them would require a communicative framework within which agents understood and gave conscious thought to what others had said. The second reason ... is that if an entity such as a computer were programmed to respond to moral demands without ever being conscious of them, then although its behavior might indeed be governed by those demands, it would not be *self-governing* in the full sense.” (Sher 2009, 69.)

Even though both of these claims seem true for Sher, he doesn't consider them strong enough to support the stringent account of the aims of moral demands. Neither of the reasons establish that each and every moral demand needs to influence the target agent's behavior through their recognition of the reason the demand provides, but at most that no moral demand can achieve its internal aim unless the target agent is conscious of not only of the demand but of *other things as well*. And thus, the conclusion seems too weak to support the view that all demands of which the agent is unaware are *ipso facto* unreasonable. And similarly, Sher adds that – *a fortiori* – it is also too weak to support the conclusion that retrospective endorsement of any demand of which the target agent is unaware is *ipso facto* unreasonable. (Sher 2009, 69.)

For Sher, it thus appears that the Kantian Principle cannot be understood in a way that would support the searchlight view – insofar as he has been able to imaginatively reconstruct an understanding of the principle that the proponents of the searchlight view might try to do themselves.

#### 3.4.2.4 *In summary: failure of the Kantian Principle to support the searchlight view*

Sher's imaginative reconstruction of the Kantian Principle, to try to fit it with the searchlight view, is admittedly very complex, and even painstakingly argued. To give a condensed outline of Sher's argument, it may be summarized in the following way:

To best try to fit the Kantian Principle with the searchlight view, so that proponents of the view could appeal to it, Sher sees that “control” and “(un)fairness” in the principle need to be understood in a very specific way. The most basic formulation of the principle being “holding agents responsible for what they did not foresee is *unfair* because their lack of foresight deprived them of *control*”, “control” needs to be taken to mean simply searchlight control, and “fairness”, all things considered, could best be understood to mean unfairness in the sense that it is in a way unreasonable to expect the target agent to be able to respond to reasons of which they are unaware or do not recognize (and, consequently, the moral demand often failing its internal aim). But even if “fairness” was understood in these terms, Sher takes the principle to not end up supporting the searchlight view, because he views that agents can be and often are responsive to (unconscious) reasons of which they are unaware. Thus, Sher concludes that the Kantian Principle fails to support the searchlight view as an adequate view. And hence it remains, for Sher, that it can be viewed as fair to hold agents responsible for failing to do what they had most reason to do, even if they were unaware that they had most reason to do it.<sup>64</sup> (Sher 2009, 55–69; Zimmerman 2009, 250–251.)

Sher sees no other live alternative routes to support the searchlight view besides his considerations about responsibility as a practical concept (3.4.1), and his considerations about responsibility and the Kantian Principle (3.4.2). That being the case, and since both considerations seem to have failed to support the searchlight view, Sher considers his case that we ought to question the searchlight view more supported than the alternative of questioning our common intuitions; he has not found any considerations that he would consider compelling us to ignore all cognitive states to which the agent lacked access to.

As illustrated earlier, Sher's overall intention seems to be to demonstrate that as the searchlight view considers *only* the conscious aspects of an agent to be relevant for moral responsibility, it cannot be a viable view (given our intuitions that Sher sees to be at odds with the view). However, in addition, he also intends to lay the groundwork for what he considers to be a more adequate interpretation of the epistemic condition – one that, tentatively:

“...instead of taking an agent's responsibility to be a simple function of what he consciously believed when he acted, we must take it to be a more complex function of his conscious beliefs on the one hand and certain objective facts about him and his situation on the other. Although the state of his consciousness when he acted is bound to remain significant, the significance of the different things of which he was aware is bound to depend on facts of which in their turn he was *not* aware.” (Sher 2009, 11.)

## 4 SHER'S ACCOUNT OF RESPONSIBILITY'S EPISTEMIC CONDITION

After arguing for the inadequacy of the searchlight view and having conducted his imaginative reconstruction, Sher moves on to formulate what he considers a more adequate alternative for how we should understand the epistemic condition of responsibility. In this chapter, I introduce and outline Sher's alternative. His formulation aims to better match with our intuitions and practices concerning responsibility, while also avoiding what he considers the searchlight view's failure to properly distinguish between the engaged and deliberative perspectives.

This chapter is divided into two primary parts to account for the two-part progression Sher himself goes through in formulating his account of the epistemic condition, and an additional third part to briefly illustrate Sher's views on the voluntariness condition. The partial account of responsibility's epistemic condition (abbr. PEC), which I introduce in section 4.1, is Sher's first, partial formulation of the condition, only dealing with *negative* responsibility. The full account of responsibility's epistemic condition (abbr. FEC), which I introduce in section 4.2, expands the initial formulation of PEC in such a way as to include also *positive* responsibility. Thus, in its final and full form, Sher presents us with an account of responsibility's epistemic condition – i.e., the knowledge requirement of moral (and prudential) responsibility – that accounts for cases of both negative and positive responsibility. Section 4.3 further addresses how his account may incorporate the voluntariness condition.

### 4.1 Partial Account of Responsibility's Epistemic Condition

I shall first use Sher's condensed formulation of the partial account of responsibility's epistemic condition (henceforth, PEC) as an introduction, and then explicate his ideas behind it. PEC is his first, partial formulation of responsibility's epistemic condition, aiming to provide what he considers a more adequate way to understand the condition as compared to the searchlight view. PEC deals solely with negative responsibility. His formulation is the following:

“When someone performs a wrong or foolish act in a way that satisfies the voluntariness condition, and when he also satisfied any other conditions for responsibility that are independent of the epistemic condition, he is responsible for his act's wrongness or foolishness if, but only if, he either

(1) is aware that the act is wrong or foolish when he performs it, or else

- (2) is unaware that the act is wrong or foolish despite having evidence for its wrongness or foolishness his failure to recognize which
- (a) falls below some applicable standard, and
  - (b) is caused by the interaction of some combination of his constitutive attitudes, dispositions, and traits.”<sup>65</sup> (Sher 2009, 88.)

In prefacing the two clauses that PEC consists of, Sher is careful to separate the freedom requirement and any other possible requirements for responsibility from the knowledge requirement, so that he can focus on the latter (see Aristotle’s original distinction, introduced in chapter 2). However, he calls the condition that can be viewed to effectively take the place of the freedom requirement as the “voluntariness condition”, and similarly the knowledge requirement is now the “epistemic condition”. “Voluntariness condition” may be seen to be a more fitting term than “freedom requirement”, as it avoids taking any implicit stances on free will. “Epistemic condition” is also in its implicit associations much more a fitting term in Sher’s context, as the condition doesn’t necessarily require awareness of knowledge (clause 1), as it may sometimes merely require a certain epistemic or cognitive context that doesn’t include awareness (clause 2).

The two clauses (or disjuncts) of PEC present the two separate conditions under which an agent can be responsible for an act that would, itself, be considered wrong (or foolish).<sup>66</sup> Clause 1 restates a version of the searchlight view. Sher agrees that the searchlight view is a necessary condition for a full account of responsibility’s epistemic condition, but not a sufficient one as it doesn’t seem to work in all situations (see Sher’s nine example cases, introduced in sect. 3.3.2). The first clause of PEC is thus, essentially, the clause that concerns what we are *aware of* while performing the act (and thus relating to the conscious, deliberative perspective). Given the wide appeal of the searchlight view, and having already presented the view earlier, Sher considers it to be a straightforward condition (Sher 2009, 89). Hence, the focus here is on the latter clause that Sher suggests as an addition.

Clause 2 is also necessary, according to Sher, so that in all situations we better not only match our intuitions about and actual practices concerning responsibility, but also so we remain consistent in our intuitive ways of thinking about responsibility (see sect. 3.3.1 & 3.3.2). This requires us to consider some aspects of the agent’s consciousness that s/he was not necessarily aware of when performing the act. The second clause of PEC is thus, essentially, the clause that concerns what we are *unaware of* while performing the act, but that can still provide basis for responsibility (and that relates to unconscious aspects of the agent).

As can be seen, clause 2 is divided into two mutually necessary sub-clauses (a) and (b). In Sher's account, these are all together required for an agent to be considered responsible when unaware of the wrongness of his or her act while performing it. Clause 2 requires the agent to have (unconscious) evidence for the wrongness of his or her act and that s/he yet fails to recognize this (i.e., is unaware of having it). The sub-clauses then specify the conditions or boundaries for the failure of recognizing the evidence. The first sub-clause, (a), requires that the failure of recognition falls below some *applicable standard* of what they should recognize. The second sub-clause, (b), additionally requires that the failure of recognition is caused by the interaction of some combination of the agent's *constitutive attitudes, dispositions, and traits*.

As also Sher himself is quick to note: The second clause of PEC is rather ambiguous or vague, in this condensed form. It requires further clarification: (i) what would an *applicable standard* for failing to recognize the evidence be in the first sub-clause of the second clause, and (ii) how would we define *constitutive attitudes, dispositions, and traits* of an agent, to fulfil the second sub-clause of the second clause, and further, (iii) why would these, together with clause 1, be an adequate explanation for *why* an agent is responsible for what they did (Sher 2009, 89 & 97). In the two upcoming section, 4.1.2 and 4.1.3, I outline Sher's answers to these questions. Section 4.1.2 focuses on (i), and 4.1.3 focuses on (ii), while both clarify (iii). In section 4.1.4, I summarize Sher's extrapolations of PEC. But before all this, in section 4.1.1, I outline how Sher tentatively demonstrates the applicability of his epistemic condition via his nine example cases.

#### 4.1.1 Application in Sher's nine example cases

As it restates a version of the searchlight view, Sher takes clause 1 to be a sufficient condition in many cases to explain our intuitions about responsibility. And considering the support that the view has gathered, he takes it to be rather uncontroversial as a necessary condition for all cases, and thus concentrates on demonstrating the applicability of PEC in cases where he considers searchlight view to fail. Namely, in cases in which agents seem intuitively responsible for wrongful acts without awareness. (Sher 2009, 89.) To tentatively demonstrate that PEC, or specifically clause 2, does justice to these kinds of cases, Sher utilizes the nine example cases (introduced in section 3.3.2).

Sher takes all the nine agents in his example cases to satisfy 2a: they all seem to fall below some (intuitive and external) applicable standard. The more demanding question he takes to be whether the agents satisfy 2b in different cases. (Sher 2009, 90.)

His example cases 5, 7, 8, and 9 – dubbed *Colicky Baby*, *Bad Joke*, *Bad Policy*, and *Bad Weather*, respectively – Sher takes to involve an agent whose character is in some aspect worse than normal. As the agents can be assumed to possess defective characters, Sher takes them to satisfy 2b especially clearly. For example, Scout failing to realize not to give the vodka to the baby, Sher takes us not to attribute to any lack of information (clause 1), but to some constitutive attitudes, dispositions, and traits of her, characterized by flawed patterns of thought (clause 2); e.g., her irresponsibility-related dispositions, perhaps her impulsiveness or the lack of consideration of consequences that prevents her from having the thought that she might be harming the baby. Similarly, the related disposition in Ryland’s case of bad joke could be, perhaps, self-absorption – and similarly the remaining cases of Sylvain and amerika. In all the cases, Sher takes us to naturally gravitate to the view that the agents themselves are the source of their failure to appreciate the connection between the facts with which they are acquainted and the wrongness of their act. (Sher 2009, 90–91.)

The four cases examined above involve agents whose characters Sher takes to be flawed. The rest of the cases Sher takes to involve agents whose characters are *not* flawed. Yet, they can still satisfy clause 2b, because in some cases the wrongness of an act can be traced to interaction of dispositions none of which are moral flaws themselves.<sup>67</sup> As immoral acts can be produced via dispositions that are not flawed, similarly unflawed dispositions can also prevent agents from responding to their *evidence* that they are acting wrongly. Thus, Sher takes it that at least in most of the remaining cases 2b applies as well. (Sher 2009, 91.)

For example, in Sher’s example case 1 – dubbed *Hot Dog* – Alessandra is not an irresponsible person, but still fails to protect her dog in a particular situation. Thus, Sher thinks, we take her failure to react to be explained by some *further* combination of her attitudes or traits, besides her disposition of generally being responsible. And even if all her attitudes, dispositions, and traits were not bad in themselves, still they were flawed or different *enough* to make her forget about the dog. For example, if Alessandra were less solicitous of her children, or if she was made less anxious by conflict, Sher thinks she would not have forgotten about the dog in the hot van. Similar line of thought can be applied to case 2, *On the Rocks*. And so, even in good-character cases – like in bad-character ones – Sher takes us to naturally gravitate to the view that the agents themselves are the source of their failure to appreciate the connection between the facts with which they are acquainted and the wrongness of their act. In other words: in both types of cases the agents’ failure is attributed to their attitudes, dispositions, and traits that make them the person they are. (Sher 2009, 91–92.)

The three remaining cases 3, 4, and 6 – *Caught off Guard*, *Home for the Holidays*, and *Jackknife*, respectively – Sher considers the least straightforward. This is because he has doubts

whether Wren's amount of care for her duty, Joliet's panic, and Father Poteet's inaccuracy of processing visual cues in traffic are themselves reflections of the complex psychology that makes the agents the person they are. Still, Sher is adamant that these characteristics would be constitutive of the agents, and to which constitutive characteristics we would intuitively associate the source of their failure. Nevertheless, even if the characteristics were not constitutive of the agents, Sher thinks it would not much undermine his case as it would still capture what he considers our intuitions about the majority of the nine counterexamples to the searchlight view. The hesitation in these cases he takes to be unsurprising, as he views these cases to be least clear in terms of whether the actions lie within the boundaries of the self, towards which responsibility would be directed. (Sher 2009, 92–93.)

Thus, Sher views that his partial formulation of the epistemic condition (PEC) successfully matches with our intuitions in most if not all examples that he considers the searchlight view to not match with. But the difficult question remains: how exactly, in a more general sense, ought we to characterize the *applicable standards*, and the criteria of what determines our *constitutive* attitudes, dispositions, and traits, in clause 2?

#### 4.1.2 Applicable standard

So, what exactly would count as an *applicable standard(s)* for a failure to recognize having evidence for an act's wrongness (in clause 2a)? In other words: what exactly would count as an applicable standard that determines what an agent in a given situation should be (or should have been) aware of?

Even though Sher takes all the nine example cases to fulfill some (external and implicit) standard of recognition, he does mention some examples that would not fulfill the standard. These examples are situations where the agent, despite their act's wrongness being possibly caused by their constitutive attitudes, dispositions, and traits, they are (intuitively) not responsible. Namely, Sher mentions the victim of a sudden heart attack (who, had they been more attuned to their body, would have avoided the situation), the teacher whose chance remark precipitates a suicide (who, had they been more psychologically insightful, would have recognized the warning signs), and the pedestrian who is swallowed by a sinkhole (who, had they had X-ray eyes, would have seen the earth opening). The question is what, specifically, are the standards separating Sher's example cases where the agents seem responsible and these latter types of cases where the agents seem not. In the former cases the agents' state of awareness – or their constitutive attitudes, dispositions, or traits – are defective against some applicable standards. (Sher 2009, 87–88.)

Sher distinguishes two challenges that he needs to solve in specific ways via two presuppositions, for the applicable standards to function the way he intends.

*Firstly*, for Sher to be able to identify the standards, it is necessary for him to distinguish (internal) facts about the agent *from* (external) facts about their situation. In other words: between the nine example cases that fulfill the standard and the abovementioned cases that do not fulfill the standard, Sher needs to distinguish between situational factors on the one hand, and the agents' constitutive attitudes, dispositions, and traits, on the other. This poses a challenge, because drawing the boundary is not immediately clear. For example, an agent's prior history, and their moral and nonmoral background beliefs, lie close to the boundary. An example situation that Sher mentions lying close to the boundary is of a homeopath who endangers his sick child's life by refusing to authorize a proven therapy (yet who theoretically has easy access to orthodox medical care and critical knowledge on the Internet).<sup>68</sup> If the agent and their situation cannot be properly distinguished – if the agent's constitutive attitudes, dispositions, and traits preventing them from realizing the moral character of their act are part *of* their situation – then their resulting lack of awareness cannot fall below any standards that apply to agents *in* a similar situation. Consequently, if the distinction was not made and thus the situation in 2a was understood in a subjective manner, the moral intuitions underlying clause 2 would be seriously undermined. (Sher 2009, 98–99.)

*Secondly*, the nature of the standards needs to be understood to be normative instead of statistical. This means that the standards – when we say someone should be aware of something, or a reasonable person in a similar situation would be aware of something – is not relying merely on what a typical or average person who found themselves in the agent's situation would realize about the moral implications of the act, but that the agent has failed to meet a demand whose force is independent of what anyone else in their situation would or would not realize. A related question is whether the normative requirement is epistemic, or moral (or prudential). The reason why Sher needs the standard to be understood as nonstatistical is because otherwise the intuitions underlying PEC would also be undermined. If the standards in 2a were merely statistical, they would be too relativized for Sher, depending on other people's reactions instead of intrinsic features of the agents' reactions. (ibid.)

Hence, to defend PEC, Sher defends the two presuppositions behind his applicable standard (in 2a): (1) a nonsubjective, specifically objective, account of the unwitting wrongdoer's situation, and (2) a nonstatistical, specifically normative, account of what one ought to be aware of. (ibid.)

Initially, Sher looks for help in legal literature, particularly in literature concerning negligence and the use of the *reasonable-person standard*.<sup>69</sup> The standard can be interpreted objectively or

subjectively, but ultimately Sher doesn't find any satisfying arguments in the legal literature for how we should interpret the standard in a moral (or prudential) context, as opposed to legal. Essentially, the legal context focuses too much on efficient administration of the law and effective control of future behavior: for example, the commonly applied nonsubjective account of the reasonable-person standard in the legal context is justified via practical, forward-looking considerations instead of moral considerations after the act. Sher is thus forced to offer his own, original justifications. (Sher 2009, 100–104.)

I examine Sher's justifications concerning his two presuppositions in the following two sections: an objective account of the unwitting wrongdoer's situation (in 4.1.2.1), and normative account of what one ought to be aware of (in 4.1.2.2). In section 4.1.2.3, I provide a condensed summary of how Sher understands clause 2a.

#### *4.1.2.1 An objective account of the boundary between the agent and the situation*

In terms of the first presupposition – an objective account of the wrongdoer's situation – Sher notes that in considering where to draw the boundary between the agent and their situation, we can equally well choose whether to examine the question in terms of what aspects of the agents should the situation include or in terms of what aspects of the situation should the agent include. These are both addressing the same boundary, but Sher thinks the latter is a more fruitful way to frame the matter in the context of epistemic condition of responsibility, as it is primarily interested in the agent rather than the situation. In other words, it is helpful to focus on where the boundary of the agent him- or herself is. This leads Sher to examine theories related to the nature of the agents themselves. (Sher 2009, 105.)

Via that body of theory, Sher supports his first presupposition, an objective account of the wrongdoer's situation. Even though Sher doesn't provide any explicit references, he notes that many philosophers, when considering what makes someone the person they are, very often appeal to some combination of the person's desires, beliefs, dispositions, and traits.<sup>70</sup> These aspects of the person are very often taken to determine the person's identity in a way that their physical attributes and social circumstances (that is, situation) do not. Albeit there is disagreement of *how much* of a person's psychology has this status, Sher considers that many philosophers would draw the line between an agent and their situation roughly to the same place as the traditional legal rule. Namely, the line would be drawn to – and the applicable standard would be understood as – *what a reasonable person in the agent's situation would realize*. (ibid.)

Thus, the more specific related question Sher wants to answer is where to locate the boundaries of *the responsible self* – i.e., the boundaries of the features of *agents* that constitute them as responsible persons in *situations* where they seem responsible without awareness. This reveals a close connection between clauses 2a and 2b: the answer is needed not only to make further sense of the applicable standard (2a), but also to answer which psychological states are constitutive of an agent (2b). (Sher 2009, 105–106.) Clarifying this connection, Sher writes:

“[F]or any given agent, the range of answers that exclude from his situation the psychological states that account for his failure to recognize his act as wrong or foolish will roughly coincide with the range of answers that construe those psychological states as constitutive of him. This suggests that the two requirements of PEC’s second disjunct—that the agent’s failure to respond to his evidence that his act is wrong or foolish be both (a) defective relative to the standards that determine what a reasonable person in his situation would realize and (b) traceable to his own constitutive attitudes and dispositions—are connected at a deep level. Although they appear distinct, the two requirements are rooted in a single theory of the responsible self.” (Sher 2009, 106.)

As a single theory of the responsible self can further clarify both 2a and 2b, Sher concludes that he doesn’t need to provide any independent arguments for the agent’s psychological states – whose interaction prevents agents from realizing the moral implications of their act – not being included in the situation. It suffices that he demonstrates that the attitudes and traits that prevent the agents from realizing the moral implications are generally constitutive of their possessors (clarifying 2b), as the conclusion that they are not elements of their possessors’ situation would immediately follow (establishing the distinction between the agent and their situation in 2a). (ibid.)

In the next section, 4.1.3, I summarize Sher’s theory of the responsible self, and thus his concluding illustration of how to understand clause 2. But first, I introduce his justification for his other presupposition concerning clause 2a.

#### 4.1.2.2 *Normative account of responsibility without awareness*

In terms of the second presupposition – a normative (instead of statistical) account of what one ought to be aware of – Sher notes that unlike the first presupposition, this is genuinely independent of how we interpret clause 2b. Thus, it requires its own defense. (Sher 2009, 106.)

Sher provides a few arguments for the normative account:

*Firstly*, he notes that even though our judgments in a given situation are often heavily influenced by our (statistical) beliefs about what *most others* in the target agent's situation *are* (or would be) aware of, a statistical interpretation of the standards would too easily be reduced to an *argumentum ad populum*. Even if the standards were ones that most people would meet, it wouldn't follow that the standards apply *because* most people meet them. (Sher 2009, 107–108.)

*Secondly*, he notes that the statistical approach – often utilized in legal context – is not applicable to moral context, because in practice it would undermine many of our moral intuitions. For example, if a virus suddenly impaired everyone's cognitive faculties in such a way that most people would forget their dogs in a hot van (like Alessandra in Sher's example case *Hot Dog*), the statistics changing would not affect our moral judgments of the matter concerning the minority who remain cognitively unimpaired. (Sher 2009, 107–108.)

*Thirdly*, even if the statistical approach was understood to be set by the *agent's own* previous level of performance – e.g., that Alessandra should have been aware of Sheba's plight because it is the sort of thing that Alessandra generally does recognize – it too would undermine many of our moral intuitions. For example, if Alessandra was just inflicted by the virus, we would not continue insisting that the impaired Alessandra was responsible unlike the unimpaired Alessandra. (Sher 2009, 109.)

Sher notes that the third argument shows that it doesn't matter what Alessandra has remembered in the past but only what she is capable of currently remembering. This suggests that instead of the statistical approach, the standards about what a person should realize at a given time may be set by that person's current cognitive capacities. Initially, for Sher, this seems like a very promising way to understand the standards. (Sher 2009, 109–110.)

However, even if the standards were understood in terms of the agent's current cognitive capacities, it is not enough: Sher notes that the standards underlying our intuitive judgments about what an agent in a particular situation should be aware of cannot be set *exclusively* by the cognitive capacities of that agent. If what Alessandra should have remembered when she was preoccupied in the school is entirely dependent on what she was capable of remembering, then she should equally remember, for example, an umbrella, and last summer's beach novel that she also noticed in the car besides Sheba. (Sher 2009, 111.)

Thus, Sher concludes that what any given agent should recognize depends not only on what that agent is capable of recognizing, but also on the agent's moral obligations; i.e., on what they must recognize if they are to discharge their moral (or prudential) *duty*. To illustrate the latter dependency: as Alessandra does not have any moral duties towards her umbrella, but does have towards Sheba, only the latter should have been recognized. This result suggests to Sher that the applicable standards of what agents should (epistemically) be aware of are rooted in the ostensibly different standards that determine what those agents morally (or prudentially) ought to do. Hence, the nature of the applicable standard, in 2a, should be understood as normative in the same way as moral "oughts" are.<sup>71</sup> (Sher 2009, 111–112.)

One possible objection to this conclusion that Sher answers is that it might be the case that morality can only give rise to standards that dictate actions – i.e., that are action-guiding – in which case the standards that determine what agents merely *should* be aware of could not rise from morality. Even if this might be the case (despite, for example, many virtue ethicists disagreeing), Sher thinks that the two could still be connected. Namely, by morality playing a role in our non-action-guiding standards by which we *assess* others. For example, *traits* like honesty and kindness are not actions, and neither are *feelings* such as pride nor *attitudes* such as indifference, but they still describe a moral character that people ought or ought not to have. Thus, Sher sees that the normativity of the applicable standard, in 2a, can at the very least be rooted in non-action-guiding moral requirements. And hence, overall, the applicable standard, in 2a, is to be understood as a joint function of (a) the agent's cognitive capacities, and (b) the [non-action-guiding] moral requirements that apply to the agent. (Sher 2009, 112–113.)

But how are (a) and (b) connected, exactly? Sher illuminates the connection as follows:

“From the premises that the demands of morality ... are directed at agents in their capacity as reason-responders and that agents cannot be reason-responders if they lack relevant cognitive capacities, we may conclude that the demands of morality ... are directed only at those who possess the relevant cognitive capacities, but not at those who lack them.” (Sher 2009, 115.)

The relevant cognitive capacities for agents to be reason-responsive (in a general sense), Sher takes to include the agents' disposition to notice various features of their surroundings, to separate what is relevant from what is not, to preserve any relevant beliefs in their short-term memory, to retrieve

information from their long-term memory as needed, and to draw the appropriate conclusions from their beliefs and goals. (Sher 2009, 114–115.)

#### 4.1.2.3 *Summary of Sher's applicable standard*

Sher's examination of the applicable standard, that he has in mind in clause 2a, to match with what he sees to be our intuitions in the nine example cases, could be summarized as follows:

The applicable standards that determine what an agent in a given situation should be (or should have been) aware of are to be understood in terms of what a reasonable person in the *agent's situation* would realize when the boundaries of *the responsible self* are properly understood (4.1.2.1), and specifically should realize in a normative sense, by relevant moral demands being directed at agents who possess cognitive capacities necessary for them to be *reason-responsive* (4.1.2.2).

Thus, further considering that the responsible self establishes a connection between clauses 2a and 2b, the agents in the example cases can be (intuitively) considered responsible without awareness in the sense that their failure of recognizing the wrongness of their act has resulted from the interaction of the attitudes, dispositions, and traits that make them the person they are.

For example, what separates Alessandra – in Sher's example case *Hot Dog* – from the victim of a sudden heart attack, is the standard that Alessandra's constitutive attitudes, dispositions, and traits were within the boundaries of the responsible self, whereas the victim of a heart attack was under a situation beyond the boundaries of the responsible self. And moral demands can be directed to Alessandra as she possesses cognitive capacities necessary for her to be reason-responsive.

To conclude his explanation of PEC, Sher sets out to draw the boundaries of the responsible self – that is, the boundary between the responsible agent and their situation. This is the topic of the following section, concluding our trek through Sher's partial account of the epistemic condition.

#### 4.1.3 *The boundaries of the responsible self*

As established in section 4.1.2.1, Sher considers a single theory of the responsible self to further clarify both clauses 2a and 2b of PEC. Via his theory of the responsible self, he aims to demonstrate that the attitudes, dispositions, and traits that prevent agents from realizing the normative moral implications of their acts are generally constitutive of their possessors (clarifying 2b), after which the conclusion that they are not elements of their possessors' situation would follow (establishing the

demarcation of the agent from their situation in 2a). Fulfilling this aim would, according to Sher, sufficiently satisfy explaining what he considers our intuitions in the counter-example cases for the searchlight view, while compactly formalizing them in PEC.

To start, Sher notes that an agent's failure to realize or remember what they should have realized or remembered can always be traced to some combination of their psychology and/or physiology. Why those subpersonal psychological or physical states should then affect our reactions to the *agent*, is due to those states being so closely related to them that it is reasonable to view the cognitive failures that arose from those states as originating *in* the agent (as oppose to their situation). In other words: the relevant subpersonal states are among the person's constitutive features. The following sections extrapolate what the subpersonal states Sher has in mind are, why he thinks they originate in the agent, and introduces the main contending views along with Sher's counterarguments for those views. (Sher 2009, 117–118.)

In section 4.1.3.1, I introduce the maximalist position and the two minimalist positions that Sher views as the main alternatives and guidelines to approach his view. In 4.1.3.2, I outline Sher's intermediate position, properly clarifying Sher's view of the constitutive features in 2b. Sections 4.1.3.3–4.1.3.5 outline Sher's defense against the first minimalist position, second minimalist position, and other intermediate positions, respectively.

#### 4.1.3.1 *The maximalist and the minimalist positions*

Sher distinguishes two extreme positions to approach how we should think of the responsible self: a maximalist position and a minimalist position. The maximalist position sees the responsible self as the whole human organism: an agent's every psychological or physical feature is equally constitutive of them. For example, their height, white blood count, and their most deeply held values are equally part of what makes them the responsible person they are. By contrast, the minimalist position sees the responsible self as constituted *strictly* by characteristics that are taken to be necessary for responsibility – such as conscious will or rationality. (Sher 2009, 118–119.)

However, Sher himself considers both the maximalist and the minimalist position as inadequate. The problem he sees with the maximalist position is that it is too inclusive. Even though humans are wholly part of the natural world, the majority of their physical (and maybe psychological) features – like their white blood count, for instance – have no obvious connection to their beliefs, judgments, decisions, actions, or omissions. In other words: many features of an agent are not in any obvious way connected to the features that make questions of responsibility meaningful. Thus, Sher

sees there to be no reason to view such features of an agent as anything that makes them responsible, or, in other words, anything that constitutes them as a responsible agent – and consequently, Sher does not address the maximalist position any further. (Sher 2009, 119.)

The problem Sher sees in the minimalist position is, conversely, that it is too exclusive. However, there are three different minimalist position candidates that Sher distinguishes and introduces.

Sher notes that there is wide agreement that any being that fully lacks consciousness, and/or is systematically unresponsive to theoretical or practical reasons, cannot be a responsible agent. Hence, consciousness and/or reason-responsiveness could more straightforwardly be argued to fall within the boundaries of the responsible self. Thus, the minimalist position could argue that the responsible self contains only (a) consciousness (i.e., the conscious center of will), (b) reason-responsiveness (i.e., ‘rationality’, rational activity, or the ability to make judgments about reasons), or (c) both. (Sher 2009, 119–120.)

The minimalist position that considers only consciousness to fall inside the boundaries of the responsible self is represented by a variant of the searchlight view that, in Sher’s view, philosopher Neil Levy has endorsed and labelled as “volitionism” (also called internalism; see sect. 5.4.3). According to this view “an agent is responsible for something (an act, omission, attitude, and so on) just in case that agent has – directly or indirectly – *chosen* that thing” (Levy 2005, 2).<sup>72</sup> And the opposite minimalist position that considers only reason-responsiveness to fall inside the boundaries of the responsible self is represented by philosophers Thomas Scanlon and Angela Smith via their view, that Sher calls “attributionism” (later also called answerability; see sect. 2.2). According to this view, agents are responsible for all of the actions, beliefs, and attitudes – conscious or not – that reflect their judgments about what they have reason to do, believe, or feel.<sup>73</sup> (Sher 2009, 120.)

To further distinguish these two minimalist positions, Sher describes volitionism – that abstracts away from everything except an agent’s conscious choice – to in effect identify the responsible self with, or in terms of, the first perspective of an agent. By contrast, attributionism – that abstracts away from everything except an agent’s judgments about reasons – in effect identifies the responsible self with an agent’s capacity to reach judgments about reasons. (Sher 2009, 119–120.)

Though Sher considers *both* minimalist positions to clearly consist of a feature connected to an agent’s responsibility, he sees them to be too exclusive even taken together. He views that consciousness and reason-responsiveness are intimately linked with certain physical and mental features. Thus, by working backwards from the minimalist features, Sher suggests we can identify a

whole array of conceptions found between the maximalist and minimalist extremes. He sees that the minimalist views require us to add one addition to them: *the causes of the minimalist features*. (Sher 2009, 121.)

#### 4.1.3.2 *Sher's intermediate position, and the constitutive features of an agent*

Sher argues that instead of the minimalist positions – of viewing the responsible self as bound by an agent's conscious will or their ability to evaluate practical and theoretical reasons and adjust their actions accordingly – we should think of the boundaries of the responsible self as consisting of an “enduring causal structure whose elements interact in ways that *give rise to* [the minimalist] responsibility-related activities”. Thus, by saying that a certain psychological or physiological state is *constitutive* of an agent, or part of what makes them the person they are, Sher means that the psychological or physiological state is “among the elements of the system whose causal interactions determine the contents of the conscious thoughts and deliberative activities in whose absence [the agent] would not qualify as responsible at all.” (Sher 2009, 121–122.)

To give a bit of a more concrete idea what kind of features of an agent Sher includes in their constitutive features, he notes that the list is quite extensive but does give some examples:

“It is a commonplace that each person's theoretical and practical decisions are influenced by factors such as his background beliefs, his moral commitments, his views about what is good and valuable, and what he notices and finds salient. His decisions are influenced, as well, by his degree of optimism or pessimism, his attitude toward risk, and many other facets of his emotional makeup. Hence, by my account, all such features of an agent will qualify as constitutive. In addition, as long as they remain compatible with the general framework of folk psychology within which the notions of agency, reasons, and responsibility are embedded, we can expect that many of the factors that explain an agent's thoughts and actions at other, deeper levels – the relevant neurophysiological mechanisms, for example, or the functionally defined constructs that populate the flow charts of cognitive psychologists – will qualify as constitutive too.” (Sher 2009, 122.)

Further, Sher emphasizes that “these claims are meant to imply not that we cannot classify someone as a responsible self without being able to identify the relevant causal structures, but only that in so classifying him, we assume that such structures do in fact exist.” (ibid.)

Although this reminds of identifying a responsible agent by their character, Sher is quick to make the distinction clear: the focus is not on the character of the agent *per se*, but the *causal nature* of the link between the minimalist features that make responsibility meaningful and the constituents of those features. (ibid.) As illustrated in section 4.1.1: even good character agents in Sher's nine example cases are considered responsible due to their cognitive mishaps being causally produced by their constitutive attitudes, disposition, and traits.

His position thus laid bare, there is one especial difficulty with it that Sher identifies. Namely: if the constitutive features of a responsible agent are ones that *sustain* the agent's rationality-related activities, then how is it possible that those same features can *prevent* the agent from engaging effectively in such activities? In other words: if the failure to appreciate an acts wrongness is caused by the agent's constitutive elements, how can those same elements warrant responsibility? (Sher 2009, 122–123.)

To answer this difficulty, Sher makes two points: one conceptual and one broadly empirical. The conceptual point is that a constitutive element that sustains an agent's rationality-related activities in general can sometimes prevent them from responding to reasons. This is not contradictory, Sher views. He considers this especially true for a complex psychological system where various constitutive features are interlinked with each other and many external factors (compare this to section 4.1.1, where Sher mentions that it is possible to trace the wrongness of an act to interaction of dispositions none of which are moral flaws themselves). (Sher 2009, 123.)

As the broadly empirical point, Sher suggests we think of a car or a computer as an analogy. A car is a whole complex system, comprised of constitutive parts like sparkplugs, for example, that occasionally do not fire – and computers are complex systems where, for example, hard drives sometimes fail. Similarly, Sher views that the constitutive features of humans may sometimes fail, causing, for example, cognitive lapses. Sher thinks it is unclear why we should not suppose that, for example, Alessandra forgetting her dog Sheba in the hot van is not constituted by aspects of her psychology that are consistent contributors to the way she characteristically approaches practical and theoretical problems (and the same with all the other agents in the nine example cases). (Sher 2009, 123–124.)

To reiterate, and to visibly and in simple terms connect Sher's intermediate position with clause 2b of PEC: It can be described that the constitutive attitudes, dispositions, and traits, that 2b requires to have caused an agent's unawareness of what they should have been aware of, are to be understood

as those causal features whose interaction brings about the agent's actions and that are required for the agent to qualify as responsible in the first place (via them giving rise to the minimalist responsibility-related activities). For example, Alessandra is responsible for forgetting Sheba because, presumably, her forgetting Sheba is brought about via her attitudes, dispositions, and traits that are not only necessary for her to be generally responsible but that also cause her to pay so much attention to untangling the situation at the school that they further cause her to forget Sheba. (see sect. 4.1.1 & 4.1.4 for more examples.)

#### 4.1.3.3 *Sher's defense of his intermediate position against volitionism, the first minimalist position*

To provide justification for his intermediate position of the boundaries of the responsible self, Sher argues that both minimalist positions, closely inspected, would actually lead to, or close to, Sher's position via their difficulties that are best resolved by augmenting them with a causal component. Further, he argues that other intermediate positions are essentially ruled out. In this and the two following sections (4.1.3.3–4.1.3.5), each of the three lines of argument are outlined.

*The first minimalist position*, to remind us, views that the responsible self is bound by consciousness. In other words, it is bound by the first-person subjectivity of the responsible agent, who is thought as a simple conscious center of will. Sher sees three problems to emerge, if we take the responsible self to be bound only by this first-person feature.

*Firstly*, Sher notes that the concept of responsibility is interpersonal. As an adequate account of the responsible self needs to make sense of both personal *and* interpersonal judgments, it needs to allow us to think of responsibility from our third-person perspective. For example, like Sher noted in the context of Korsgaard and Bok (in sect. 3.4.1), even responsibility as practical reason allows for this. First-person perspective can still play a role by us noting an agent's perspective via our third-person perspective, but responsibility cannot require us to *occupy* that perspective. Thus, an account of the responsible self bound strictly to an agent's subjectivity cannot be viable as it would not seem to allow interpersonal judgments. However, if we take the responsible agent to consist of the *causes* of the agent's subjectivity, we do not need to occupy it to make sense of interpersonal judgments: we can hold reasonable beliefs about the existence and contents of an agent's subjective consciousness via noting the sustaining causes. (Sher 2009, 124–125.)

*Secondly*, and relating to the first problem, Sher notes that the first-person account of the responsible self does not locate responsible selves *in the world*. As we cannot observe subjective consciousness, strictly speaking, our account of the responsible self would be nonempirical. But if

we take responsible selves to be bound by causally effective psychological structures, we can, at least in principle, be responsive concerning the empirical nature of the causally operative states of consciousness. (Sher 2009, 125.)

*Thirdly*, Sher notes that we cannot even view ourselves as simple conscious centers of will, even, as our consciousness is invariably restricted in the face of unconscious features of ourselves. For example, the driver (briefly mentioned in sect. 3.1) who is either actively or passively aware of many things on the road and in the car – concentrating on noticing his exit, a car approaching on his left, a passenger telling an anecdote etc. – is still only aware of the tiniest fraction of the information he needs to perform in the situation. He is not thinking about, for instance, the mechanical operations he is performing let alone what he will do in case of a flat tire or a traffic jam. Instead, he simply trusts that he will think of what is needed when the time comes. Similarly, like the driver, we all have confidence in our subjective consciousness to perform under situations we are not consciously preparing for. Thus, as this kind of reliance on cognitive features we are not conscious of seems pervasive in our practical deliberation – and our theoretical ratiocination, Sher adds – Sher views that there must be a substantial *non-conscious* dimension to the concept of ourselves that informs our deliberation and ratiocination. And thus, we cannot be coherently thought of as conscious centers of will. (Sher 2009, 125–127.)

Sher further outlines his third point as follows: under the first minimalist position, for it to be able to project our practical deliberations into the future, the deliberating self needs to be construed as “(a) containing a substantial non-conscious component, and (b) playing a causal role in generating the thoughts that rise to consciousness at different moments, and (c) generally (though not invariably) functioning in ways that sustain our reason-related activities.” This picture, Sher sees to be compatible with if not suggestive of the view he is defending. Thus, even if Sher was to accept that the responsible self is best understood via the first minimalist position, this would still provide reasons to reject viewing responsible selves as simply conscious centers of will. (Sher 2009, 127–128.)

#### *4.1.3.4 Sher’s defense of his intermediate position against attributionism; the second minimalist position*

*The second minimalist position*, to remind us, views that the responsible self is bound by reason-responsiveness. In other words, it is bound by an agent’s rationality, in the sense of their ability to form judgments about reasons.

In the context of defending his intermediate position, Sher extrapolates this minimalist position a bit more via the attributionist views of Thomas Scanlon and Angela Smith. The main takeaway of that extrapolation seems to be that, more specifically, the view – in Sher’s reading – sees agents responsible for attitudes, actions, and omissions that reflect their judgments about what they have reason to believe or do.<sup>74</sup> For example, in this view, Alessandra would be responsible for leaving Sheba in the van insofar as it results from her judgments, whether chosen consciously or not. In other words, if Alessandra can be asked to *defend* her leaving Sheba in the van, and she would be sincerely willing to do so, the act thus reflecting her evaluative judgments about reasons, she could be considered responsible. Otherwise, she would not be responsible.<sup>75</sup> (Sher 2009, 128–129.)

While attributionists do not explicitly mention the responsible self, Sher interprets them to view that the responsible self would be bound by an agent’s ability to make judgments about reasons, i.e. their reason-responsiveness. After all, that ability is what would be minimally required for an agent to constitute as a responsible agent, and contents of a person’s judgments would be what makes them the particular responsible agent they are. (Sher 2009, 129.)

Sher sees two problems to emerge, if we take the responsible self to be bound only by an agent’s ability to make rational judgments. To be precise, he does not disagree with the view *per se*, but seems to view it as insufficient.

*Firstly*, Sher notes that agents often seem responsible for acts they have made no judgments about. For example, even though Alessandra may judge her to have good reason to stay at the school and hash things out with the authorities, the act that seems wrong – leaving Sheba in the hot van – plays no part in her judgment, as she has simply forgotten about the dog. Thus, it is unclear to Sher how the wrongness of an act can be connected with Alessandra’s judgments, if the wrong-making features of the act do not even enter Alessandra’s judgments. (Sher 2009, 130.)

Sher notes that Smith seems to answer this problem by hedgingly implying that an agent forgetting about someone provides *indication* of the agent dismissing an evaluative judgment of that someone being a significant source of reasons. Sher sees no reason for disagreement with this particular claim. However, because of the hedged nature of the claim, Sher sees it to allow for failures to notice morally relevant features of acts that are *not* themselves judgment-based. For example, Alessandra does not seem to make such an evaluative judgment, as the urgency of her dispute and the high emotional tax of the situation seem to sufficiently explain her forgetting her act’s wrong-making feature; namely, forgetting Sheba in the hot van. (Sher 2009, 130–131.)

To solve this problem, Sher suggests that “we must locate the significance of Alessandra’s failure to remember Sheba not in what it reveals about her judgments about reasons, but rather in its being caused by the same psychophysical structure that sustains her ability to *make* such judgments.” In other words, to solve the problem we need to incorporate a causal component to the view. (Sher 2009, 131.) Sher doesn’t make it very explicit, but this would seem to solve the problem in the sense that Alessandra could still be held responsible in a (comparatively) straightforward manner when the responsible self is located in the causal structure – her constitutive attitudes, dispositions, and traits – that results in her forgetting about Sheba in the hot van, whereas with Smith’s view, agents like Alessandra could be considered not responsible despite their act (intuitively) warranting responsibility.

*Secondly*, Sher notes that attributionism, in its basic form, is too straightforward: we can imagine agents whose actions, beliefs, and feelings are grounded in their judgments about what they have reason to do, believe, and feel, but who we would *not* view responsible. He gives a few examples of agents whose reasons seem to be detached from reality. For example, “someone might take the fact that it is raining to be a reason to shave the left side of their head, the fact that the floor is not swept to be a reason to expect a financial windfall, and the fact that his cat’s nose is running to be grounds for intense envy”. Given that these beliefs are also not backed by any less implausible beliefs, we would not seem to consider such agents responsible for their beliefs. Thus, as the basic version of attributionism seems to be detached from *actual* reasons for doing something, it would need to incorporate this into its view of the responsible self. Namely, the responsible self would be expanded from an agent’s ability to form judgments about reasons to the more demanding agent’s ability to form *accurate* judgments about reasons. (Sher 2009, 131–132.)

Unless the attributionist somehow resists this addition, Sher considers them to be well on their way to introducing a causal element into their account. This further addition Sher views likely to be necessary because it is hard to see how an agent’s judgments about his or her reasons could systematically be accurate without being causally dependent on their *truth-making* features. Thus, the causal features on which the agent’s accurate reason-responsiveness depends on would seem to be just as constitutive of them as accurate reason-responsiveness itself. And thus, attributionism, like volitionism before, would seem to lead close to Sher’s own intermediate account of the responsible self, which incorporates the causal aspects of the minimalist features into the account. (Sher 2009, 132–133.)

#### 4.1.3.5 *Sher's defense of his intermediate position against other intermediate positions*

Sher acknowledges that as there is a lot of room between the maximalist position and the two minimalist positions, his account is hardly the only intermediate position. Thus, he examines some alternatives which also take responsible agents to be constituted by some but not all elements of the agents' psychology, but who disagree on what those elements are. Namely, he distinguishes two broad groups under which the most important alternatives fall. The first approaches the matter synchronically, i.e. in terms of what currently is the case for an agent, and the second approaches the matter diachronically, i.e. how an agent has developed through time. (2009, 133–134.)

The first, synchronic group, more specifically, views responsible agents to be constituted by only aspects of their character that they in some sense accept. The acceptance is understood in terms of an agent "wanting to have and be moved by one's lower-order desires, identifying with certain attitudes while dissociating oneself from others, and being unable to avoid these forms of identification or dissociation." Sher describes this view to be inspired by the work of philosopher Harry Frankfurt.<sup>76</sup> (Sher 2009, 134.)

The second, diachronic group, more specifically, views responsible agents to be constituted by only aspects of their character that have the right kind of causal history. This view aims to eliminate attitudes with objectionable causes from the agent's constitutive features. For example, if an attitude is caused by manipulation, brainwashing, or certain forms of conditioning, it would not be considered constitutive.<sup>77</sup> (ibid.)

Sher further notes that these two groups can be combined. We may at the same time hold that an agent is constituted only by attitudes they accept, while eliminating attitudes with objectionable causes. (Sher 2009, 135.)

However, Sher emphasizes that neither one nor both accounts can be combined with his account. To refresh our memory, Sher's view asserts that a responsible agent is constituted by their full set of psychological states whose interaction sustains their consciousness and rationality-related activities. This is incompatible with an agent being constituted by their attitudes they accept because by combining that account with Sher's, a person could have one and the same psychological state that both sustains their rationality-related activities, and was thus constitutive of them, but which would at the same time be an aspect of their personality that the agent rejects, and thus *not* constitutive of them. As this would be a contradiction, the views would be incompatible. Furthermore, an agent's rationality-related activities, or the features that sustain them, might also depend on features an agent might want to reject. Likewise, trying to combine Sher's view with the view of an agent being

constituted by characters with the right kind of causal history would lead to a similar contradiction. (ibid.)

These views, even though incompatible with Sher's, do not much bother him. This is because he views them to be weakly motivated. In the case of the 'acceptance' view, Sher merely briefly mentions that insofar as an agent may have hard time to accept some aspects of their personality – whether they view them as alien or external or other – it would seem to be just as accurate to say that they are appalled because *the person they are* is that way. Further, he mentions that someone having strong feelings about some aspect of their personality being alien or external does not seem like a good basis to conclude that is really the case.<sup>78</sup> And in the case of the 'right-kind-of-causal-history' view, Sher briefly mentions that someone having been brainwashed or indoctrinated, for example, seem to render them outside of responsibility not because of their history but because they have been rendered incapable of responding to certain kinds of reasons, which has little to do with their history and a lot to do with their current psychological states.<sup>79</sup> In any case, Sher emphasizes that consciousness and reason-responsiveness seem to be minimally required for responsibility, which roughly leads to the causal view he endorses, and consequently (and even without his causal additions) rule out these competing intermediate views. (Sher 2009, 135–136.)

#### 4.1.4 Summary of Sher's extrapolation of PEC

As Sher set out to provide further illumination to both clauses 2a and 2b of PEC, it is useful to remind ourselves of what those clauses are. Presented fully in the beginning of section 4.1, clause 2 stipulates what is required for an agent to be responsible without awareness. Namely, it requires that the agent's act is wrong despite them having evidence for its wrongness and that this failure of recognition (a) falls below an *applicable standard*, and (b) is caused by the interaction of some combination of the agent's *constitutive attitudes, dispositions, and traits*. Sher carefully examined both clauses. Below is the condensed summary of that examination, more thoroughly expressed in the previous sections.

Section 4.1.2 initially clarified the meaning of the applicable standard(s), which section 4.1.3 further clarified via Sher's theory of the responsible self while also clarifying what Sher means by the agent's constitutive attitudes, dispositions, and traits. To put it bluntly: The *applicable standard* refers to what a reasonable person in the agent's situation *should* realize (or should have realized) when the boundary between the person and their situation is understood via the boundaries of the responsible self. Namely, the boundary is defined to be drawn to the agent's *constitutive attitudes, dispositions, and traits* that are defined in terms of being features of the agent – many of them

unconscious features – that are causally required for him/her to be conscious and reason-responsive; i.e. required for the agent to possess cognitive features that are necessarily required for them to qualify as a responsible agent and further as the particular responsible agent they are. And the applicable standard – the “should” proposed by it – is to be understood as a normative demand, grounded in moral demands directed at agents who possess cognitive capacities necessary for them to be reason-responsive. This Sher views to provide us an account of responsibility without awareness and an answer to how we can formalize, to clause 2, what he considers our intuitions in the nine example cases.

Sher’s case is thoroughly argued, him noting all the major counterpositions: most notably the maximalist, minimalist, and other intermediate views of the responsible self (see 4.1.3). Still, his case is quite complicated, especially when briefly summarized. Thus, it is best illustrated via examples.

Let’s think about Alessandra, for example, who forgot Sheba in the hot van due to her being distracted by administrative bungling at her children’s school, and who, presumably, could have avoided the situation had she been less solicitous of her children or if she was made less anxious by conflict. Further, let’s think about the man – and let’s say we call him Daniel – who had a heart attack, and who, presumably, could have avoided it had he been more attuned to his body.<sup>80</sup> Sher considers Alessandra responsible without awareness for forgetting Sheba, but he doesn’t consider Daniel responsible without awareness for having a heart attack, even though both could have avoided the situation had something occurred to them when it did not. Both are considered reason-responsive. The difference between these cases illustrates Sher’s thoughts: the difference is that he considers Alessandra’s constitutive properties to include being (too) solicitous of her children and being made (too) anxious by conflict, for example, whereas he considers Daniel’s happenstance to *not* be traced to properties of Daniel that are constitutive of him. In other words, Sher might view that Alessandra forgetting Sheba was due to features of her cognition that are elements of the psychological causal structure that guides her thoughts and deliberative activities and sustains her as a responsible person; whereas Daniel’s heart attack is not traceable to the psychological causal structure that guides his thoughts and deliberative activities and sustains him as a responsible person. The same distinction applies, in Sher’s view, to all the nine example agents on the one hand, who are responsible, and on the other, for example, to the teacher whose chance remark precipitates a suicide, who is *not* responsible. The suicide is not traceable to the teacher’s constitutive attitudes, dispositions, and traits; but, presumably, in all the nine example cases the agents’ acts are traceable to the respective constituents of the agents. (see 4.1.1 for more examples.)

Via a simplified summary, Wieland (2017, 17) further illustrates Alessandra to qualify as responsible, in Sher's account, because (1) her failure falls below a reasonable standard in the sense that dog owners are expected to remember their dog when they are sitting vulnerable in a hot car, and because (2) her failure is caused by her constitutive features in the sense that it is traceable to her unfortunately weak disposition to keep important things in her mind.

Sher further aimed to provide a justification for why he considers his account to be an adequate explanation for why an agent is responsible for what they did. To be clear, his account so far has only dealt with negative responsibility. But insofar as negative responsibility is concerned, Sher would seem to view that his account is adequate due to its seemingly robust nature against all the major counterpositions he considered: searchlight view he argued to be not only incompatible with our intuitions in some cases (see 3.3.2) but also unsuccessfully defended via his imaginative reconstruction (see 3.4), while incorporating it to PEC as clause 1 for cases it is compatible with; and as for his clause 2, he argued widely against major counterpositions (see 4.1.1–4.1.3). However, to complete his account of the epistemic condition, Sher also wants to incorporate cases of positive responsibility to it, via his *full* account of responsibility's epistemic condition, which we'll examine in section 4.2.

To briefly reiterate Sher's position in a bit different terms still, philosopher Bruce N. Waller (2014, 642) provides an illuminating, albeit perhaps a bit reconstructed take on Sher's position in his review. He characterizes Sher's main argument, overall, being that the constitutive deliberative elements that make us morally responsible by the searchlight view standard are the *same* constitutive factors operating when we act *without* full conscious awareness of some relevant facts that we possess unconsciously (e.g., Alessandra possessing the fact about Sheba being in the car before forgetting it due to her presumably constitutive unconscious features). Therefore, since we are responsible in searchlight cases – i.e., in cases where we know we are in the wrong (clause 1 of PEC) – and since the same mechanism is at work also in cases where our intuitions say an agent is responsible without fitting the searchlight cases (see sect. 3.3.2), for the sake of consistency we need to hold people responsible also in those latter cases (clause 2).

Even though Sher himself doesn't much explicitly discuss contemporary psychology, Waller (2014, 639–641) also notes how Sher's examination is relevant in that context. He notes that psychological research reveals how we are constantly influenced by our situational factors, cognitive biases, and other such properties that are largely inaccessible to our conscious awareness. Despite what our day-to-day conscious experience may suggest, the thought processes behind our awareness

are heavily influenced – and often controlled – by non-conscious factors. This includes any deliberative processes we engage in: in many cases we, as conscious deliberators, merely offer rationalizations for decisions or judgments made *a priori* on a deeper non-conscious level.<sup>81</sup> In the spirit of Sher, one may thus ask: why should merely the part of us who is aware be responsible, when the thought processes that we are aware of rely on processes we are not aware of?

## 4.2 Full Account of Responsibility's Epistemic Condition

As PEC only deals with negative responsibility, Sher wants to expand it to the full account of responsibility's epistemic condition (henceforth, FEC) to also account for positive responsibility. FEC is thus Sher's complete formulation of the epistemic condition. Much like in the previous chapter, I shall first introduce Sher's formulation of FEC as an introduction, and then explicate what he has added to PEC so that positive responsibility would also be accounted for in our understanding of the epistemic condition of responsibility. Due to the comparable simplicity of positive responsibility, and my overall focus being on negative responsibility, this task of illustrating the additions for positive responsibility in FEC will be relatively less demanding than illustrating PEC. The formulation of FEC is the following:

“When someone performs an act in a way that satisfies the voluntariness condition, and when he also satisfies any other conditions for responsibility that are independent of the epistemic condition, he is responsible for his act's morally or prudentially relevant feature if, but only if, he either

- (1) is consciously aware that the act has that feature (i.e., is wrong or foolish or right or prudent) when he performs it; or else
- (2) is unaware that the act is wrong or foolish despite having evidence for its wrongness or foolishness his failure to recognize which
  - (a) falls below some applicable standard, and
  - (b) is caused by the interaction of some combination of his constitutive attitudes, dispositions, and traits; or else
- (3) is unaware that the act is right or prudent despite having made enough cognitive contact with the evidence for its rightness or prudence to enable him to perform the act on that basis.” (Sher 2009, 143.)

Compared to PEC (sect. 4.1), in FEC clause 1 now includes a version of the searchlight view that applies to both positive responsibility as well as negative, where the agent was aware of the act having that feature. Clause 2 remains identical to what it was, applying solely to negative responsibility. And clause 3 is an addition intended to account for positive responsibility in cases where the agent was unaware of the act being morally right, but in which cases our intuitions would still deem them positively responsible. (Sher 2009, 142–143.)

Notably, all cases where the agent is *aware* of the act having the morally relevant feature, positive or negative, are fit into singular clause 1. But the cases where the agent is *unaware* of the act having the morally relevant feature, positive or negative, but would still be deemed responsible, requires two clauses – 2 and 3. This discrepancy, and the need for a third clause, is explained by Sher via the “asymmetry between our judgments about positive and negative cases”. The asymmetry is demonstrated via our judgments in two types of cases: (1) cases where all other things being equal except the outcome (positive vs. negative), we would (intuitively) consider an agent negatively responsible for a negative outcome, yet *not* positively responsible for a positive outcome, and (2) cases where agents are considered positively responsible without awareness. (Sher 2009, 137–141.)

To illustrate the asymmetry between cases of positive versus negative outcomes, Sher presents the following positive variation of his negative case *Colicky Baby*, where, in the original case, Scout gave vodka to the baby (the full original case introduced in sect. 3.3.2):

“*Periorbital Cellulitis*. Scout is again in charge of her sister’s colicky baby, and the baby has again been crying for hours. In this variant, though, the baby also has periorbital cellulitis, a painless eye infection that can travel to the brain if not treated quickly. Although the baby’s eye is grotesquely swollen, Scout does not register that anything is amiss. However, because she finds the baby’s fussing intensely irritating, she takes it to a nearby Urgent Care Center to get it calmed down. There the periorbital cellulitis is noticed and treated and the baby’s life is saved.” (Sher 2009, 138.)

To make the comparison clear, Sher encourages us to suppose that Scout’s evidence that the baby is seriously ill in *Periorbital Cellulitis* is just as strong as her evidence that giving the baby vodka will make it sick in *Colicky Baby*. Also, he encourages us to suppose that Scout’s failure in both cases to process her evidence is just as substandard, and equally due to her constitutive psychology. And finally, Sher encourages us to suppose that in *Periorbital Cellulitis*, Scout dislikes the baby so much that she would *not* bring it to the Urgent Care Center if she *did* realize its eye condition was life-threatening. These suppositions are intended to make the two cases as close to each other as possible,

in their relevant features. In both cases Scout's constitutive psychology prevents her from recognizing the morally relevant aspects of her situation, but Sher considers that we would not (intuitively) hold her positively responsible in *Periorbital Cellulitis*, even though we would (intuitively) hold her negatively responsible in the parallel but negative *Colicky Baby*. (Sher 2009, 138.)

To further illustrate how he views the asymmetry, Sher notes that there is a structural difference between the cases, rooted in morality's action-guiding nature.<sup>82</sup> Morality's function is two-fold: to classify acts as right or wrong, and to give us *reasons* to perform some acts and to avoid others. Thus, to function morally, we need not only do what is right, but also do it *because* it is right. Consequently, when we either do what is wrong or when we do what is right but for wrong reasons, we are not acting morally. Thus, in acting morally, there are two ways of getting it wrong, and only one way of getting it right. Following this, in *Colicky Baby*, Scout's constitutive psychology prevents her from seeing the wrong-making feature of her act, and thus, in Sher's view, she is not acting morally in the sense that she is acting wrongly. And in *Periorbital Cellulitis*, Scout's constitutive psychology prevents her from seeing the right-making feature of her act, and thus, in Sher's view, she is not acting morally in the sense that she is acting rightly for the wrong reasons (she is also not acting immorally, but ambiguously). (Sher 2009, 139–141.)

As Scout in *Periorbital Cellulitis* is acting ambiguously, thus responsibility being not applicable, yet in *Colicky Baby* immorally, thus responsibility being applicable, a full account of responsibility's epistemic condition needs to account for the asymmetry. Sher accounts for this by limiting clause 2 strictly to account for what he considers our intuitions in negative cases. Clause 1 – representing the searchlight view – he expands to apply to both negative and positive cases, because if an agent is aware of their act having a positive feature or a negative, they are equally informed to satisfy the epistemic condition. This discrepancy between clause 1 and clause 2 limits Scout in *Periorbital Cellulitis* out of the epistemic condition. (Sher 2009, 141–142.)

However, clause 3 is still needed as there are cases where Sher considers an agent being (intuitively) *positively* responsible without awareness. These sorts of agents are doing the (normatively) right act without being aware of either what makes their act right or what right reasons they have for acting the way they act. Insofar as agents are unconsciously in touch enough with the act's rightness and the right reasons for acting, Sher views them to be positively responsible. As an example, he mentions Mark Twain's Huckleberry Finn, who seems positively responsible for resisting returning Jim to his owner due to unconsciously responding to the demands of friendship or the evils of slavery.<sup>83</sup> Thus, Sher's clause 3 essentially states that agents can be positively responsibly

by accurately but unconsciously processing the information they have access to. This completes Sher's full account of responsibility's epistemic condition. (Sher 2009, 142–143.)

The strength of the final formulation, FEC, according to Sher, is that although it is “complicated and unlovely”, and its last clause is “far from precise”, it does capture “the full range of our intuitions, establishing a suitable epistemic link between an agent and each type of act for which he can be responsible, and accommodating the natural explanation of the asymmetry between our judgments about positive and negative cases.” Sher is adamant that “any adequate alternative to the searchlight view is likely to take roughly this form.” (Sher 2009, 144.)

### 4.3 The Remaining Conundrum of the Voluntariness Condition

One noticeable aspect of Sher's account that seems to differ in relation to other, perhaps more traditional ways of understanding responsibility, seems to be its relationship with the freedom requirement. As we can recall, freedom requirement was Aristotle's other classic necessary requirement for responsibility, in addition to the knowledge requirement or the epistemic condition (see ch. 2). Sher calls the requirement *voluntariness condition*. Oftentimes, the condition is also seen roughly synonymous to control condition or control requirement. However, as Sher himself illustrates, FEC forces us to re-evaluate, and ultimately abandon *control* as a requirement for responsibility. In this section, I outline why this is and what Sher proposes as the guidelines for developing a voluntariness condition *without* control nor awareness, and that would thus conform with FEC.

FEC's clauses 2 and 3 – i.e., responsibility without awareness – conflict with control in the sense that if an agent is unaware of their act's wrong-making features (in clause 2) or right-making features (in clause 3), then the agent is not consciously in control of choosing to perform the act despite its wrongness or because it is right. In other words, when an agent is acting without awareness, they are not in control of their act. On the other hand, clause 1 – i.e., the searchlight view – does not conflict with control as agents are considered only responsible with awareness. Thus, if voluntariness condition is taken to include the control requirement, also including the requirement of awareness, it would undermine Sher's account. (Sher 2009, 144–146.)

The most straightforward solution would be to abandon the voluntariness condition altogether. However, this is not an option for Sher, because he is attempting to account for what he considers our intuitions, and notes that the control requirement enjoys wide appeal. The solution he suggests is to

examine premises behind the control requirement and find if some of them can be dismissed while others accepted in a way that both preserves our deeper intuitions while also fitting together with his epistemic condition. Accordingly, the voluntariness condition would not include the control requirement but only those premises behind it that fit together with FEC. (Sher 2009, 146.)

To illustrate a key premise behind the control requirement, Sher considers what he calls the *origination* relation. As we direct reactions to agents that are justified (if they are) by the morally relevant features of their act (e.g., wrongness or rightness), the reactions must be grounded to some relation between an agent and their act's morally relevant features. Sher considers that relation, grounding our reactions, to be that the morally relevant features originate from the agent (that is, from their constitutive features). Thus, an agent's responsibility is restricted to those features of an act to which they stand in the origination relation. *As it is*, this requirement is very abstract. However, it becomes the control requirement when it is added that to stand in the origination relation to a given feature of a given act, the agent needs to choose to perform the act while being fully aware of it having that feature. (Sher 2009, 147; see also 2006, ch. 2 & 3.)

Sher considers the abstract requirement to lurk behind the control requirement. Further, he considers it to be "forced upon us by a deep structural fact about responsibility", whereas the control requirement is not needed to make sense of it. It may be that it could be defended that it is needed, but until such a defense emerges, Sher proposes that the abstract requirement is instead understood in a way that does not require awareness. He justifies this suggestion via referring to his defense of why there would be responsibility without awareness in FEC, namely the failures to justify the searchlight view (see sect. 3.4). Consequently, Sher considers the searchlight view and the control requirement to be best regarded as different aspects of a single theory. Thus, Sher's argument against the searchlight view is also an argument against the control requirement. (Sher 2009, 147–149.)

Sher still emphasizes that an agent who was fully aware of events that were occurring but was in no sense the voluntary author of those events would not be responsible for any aspect of them (Sher 2009, 148–149). So, what might a voluntariness condition that requires no awareness nor control look like?

The question is left quite open, but Sher points towards possible answers in the literature concerning free will. Namely, he points to four compatibilist accounts and one incompatibilist account that do not *prima facie* necessitate awareness as a condition for voluntariness. The compatibilist accounts include: (1) an account that lists factors standardly taken to deprive agents of

freedom – e.g., compulsion, coercion, insanity etc. – and classifies them outside of the criteria for a voluntary action<sup>84</sup> ; (2) an account that takes an act to be praiseworthy or blameworthy, and hence voluntary in the relevant sense, whenever the act’s moral quality reflects the agent’s character<sup>85</sup> ; (3) an account where an agent is taken to be voluntary if they are responsive to good reasons<sup>86</sup> ; and (4) a variant of the previous in which an act is taken to be voluntary if the mechanism that produces the action was itself reason-responsive<sup>87</sup>. Sher asserts that even though none of these accounts compel us to detach awareness from an act’s voluntariness, they can each be developed in a way that would accomplish this. The one incompatibilist account, that Sher mentions, is Robert Kane’s version of libertarianism. Kane (1998) views that an agent’s will is not free unless the antecedents of his or her act include uncaused microphysical events within their brain, of which the agent themselves is unaware even though they phenomenologically experience the events as “efforts of will”. The ultimate defensibility of all these accounts remains to be seen, but Sher is confident that a suitable account would be found within these alternatives. (Sher 2009, 149–151.)

I suspect Sher’s intuitions would encourage him to formulate a compatibilist account of the voluntariness condition. In such an account, voluntariness might be understood via factors of agents that can be traced back to their constitutive features – many of them unconscious – that enable them to be reason-responsive (see the four compatibilist accounts above). This would be very much in line with his account of the epistemic condition (see sect. 4.1 & 4.2). For example, Alessandra – who left her dog Sheba in the hot van (see sect. 3.3.2) – would not have been responsible for the act had someone or something (external) physically prevented her from attending to Sheba, thus her being involuntarily forced to be unable to fulfil her duty. This would allow any agent who is fully aware of what is occurring but who is not the voluntary author of those events to be pardonable, while not requiring control. Thus, for an agent to satisfy the voluntariness condition, it would seem to suffice that they are ‘free’ to comply with the demands of their duty, with or without awareness nor control.

## 5 RESPONSES TO SHER

George Sher's examination of the epistemic condition for moral responsibility seems to have received a considerable amount of positive feedback in the academic philosophical community. In many places, his book is viewed as a useful project that brings light to an often-overlooked element in responsibility. However, at the same time, his argument or conclusions have in many places been questioned. His book has generated a lot of discussion, essentially seeming to have catalyzed the emerging interest in the epistemic condition. (e.g., Nelkin 2011b; A. M. Smith 2010; Talbert 2011; Waller 2014; Zimmerman 2009.)

In this chapter, I go through some of the feedback Sher's examination has received. In section 5.1, I introduce a counterposition by Michael J. Zimmerman, who is critical of Sher's account, and in favor of a *qualified* searchlight view or a *revisionist* account of the epistemic condition. In section 5.2, I introduce Angela Smith's views on Sher's account. In section 5.3, I introduce some remarks that have been made by various writers in their reviews of Sher's book. And in section 5.4, I summarize some of the most recent discussion and further alternatives to how to understand the epistemic condition.

### 5.1 Michael J. Zimmerman's Argument for a Qualified Searchlight View

Sher mentions the contemporary philosopher Michael J. Zimmerman's views as the second indirect piece of evidence for the appeal of the searchlight view among academic philosophers, as introduced in section 3.2 (Sher 2009, 8–9). It is therefore fitting that Zimmerman (2009) should review Sher's book, while, in the process of answering to Sher, making his views more explicit. His views provide a good point of comparison to Sher's – and, as it happens, to other views as well (see sect. 5.4).

In section 5.1.1, I outline Zimmerman's summary of four distinct views that Sher argued against, including Zimmerman's own view. In 5.1.2, Zimmerman's view is outlined, resulting into what he calls the Origination Thesis, which is later (in sect. 5.4) revealed to be central to the contemporary discussion about the epistemic condition. 5.1.3 outlines Zimmerman's suspicions that there might be different kinds of moral evaluability, which may allow some of the disagreements to be solved. In 5.1.4, I conclude with my own initial thoughts about Zimmerman's views as compared to Sher's.

### 5.1.1 Examination of the views Sher rejects

As I introduced in section 3.3.2, Sher presents nine example cases to support his assertion that there are cases where our intuitions do not match with the searchlight view. Zimmerman, while being critical of Sher's overall criticism of the searchlight view, agrees with his statement that in many of the example cases "the agent would definitely be blamed and might well be liable to punishment" (Sher 2009, 24; Zimmerman 2009, 250 & 260). Zimmerman states that this is an accurate description of the sort of reaction that the agent can expect. Furthermore, he writes: "Sher's account of the epistemic condition of responsibility would appear to capture [our everyday moral judgments] well, certainly better than any other account that I know of" (Zimmerman 2009, 254). Thus, contrary to the searchlight view, it seems that according to *both* Zimmerman and Sher, our common intuitions would indeed hold people responsible in the kinds of situations demonstrated in Sher's example cases.

As mentioned in section 3.3.2, the question then becomes: Should we try to adjust our intuitions to better match with the searchlight view or something similar, or should we try to accommodate these kinds of situations within our thesis concerning the role of awareness, and knowledge, in moral responsibility?

The question is decided by which of the two options can be better justified via a stronger case. As presented in chapters 3 and 4, Sher argues for accommodating our common intuitions within our thesis concerning the epistemic condition of responsibility. But Zimmerman, on the other hand, argues more so for the former option: namely, that we should be more critical of our intuitions. He thinks many everyday intuitive judgments of the general populace are misguided. (Zimmerman 2009, 254.)

Zimmerman neatly summarizes most if not all the views Sher argues against. These are specifically regarding negative responsibility, which he focuses on, and which is also our primary focus (quoted directly from Zimmerman 2009, 252):

- (A) Someone is negatively responsible for some act if and only if
  - (1) he satisfies any non-epistemic condition necessary for such responsibility; and
  - (2) he was aware of the act's wrongness or foolishness.
- (B) Someone is negatively responsible for some act if and only if
  - (1) he satisfies any non-epistemic condition necessary for such responsibility; and

- (2) either (a) he was aware of the act's wrongness or foolishness or (b) he was unaware of the act's wrongness or foolishness and this unawareness was a consequence of some other act of whose wrongness or foolishness he was aware.
- (C) Someone is negatively responsible for some act if and only if
  - (1) he satisfies any non-epistemic condition necessary for such responsibility; and
  - (2) either (a) he was aware of the act's wrongness or foolishness or (b) he should have been aware of the act's wrongness or foolishness.

Thesis A, which Zimmerman calls the *unqualified* version of the searchlight view, is the most basic formulation of the view. Both Sher and Zimmerman reject this view. As illustrated in chapter 3, Sher rejects it because it doesn't account for responsibility without awareness, that he sees necessary to include in our account of the epistemic condition due to what he considers our common intuitions about responsibility. In this way, Sher sees Thesis A as incomplete, but doesn't dismiss it wholly. Likewise, Zimmerman also rejects Thesis A but not only because it doesn't seem to accommodate the full range of our everyday judgments of responsibility, but also because it doesn't account the arguments that have convinced Zimmerman of the trueness of the *qualified* searchlight view, i.e. the trueness of Thesis B (Zimmerman 2009, 254).

Thesis C could be characterized, like Zimmerman does, as the "*knew or should have known*" thesis (Zimmerman 2009, 254). This is also rejected by both Zimmerman and Sher, while Zimmerman also adds that he believes it to be a thesis to which many people appeal, whether explicitly or implicitly, when assigning responsibility in Sher's example cases. Both Sher and Zimmerman seem to agree on the basic reason for why Thesis C should be rejected. Namely, as Zimmerman summarizes what he also considers Sher to say, in effect:

"[T]hose who appeal to [Thesis C] acknowledge that one can have an excuse for wrongdoing and hence that wrongdoing does not suffice for blameworthiness. One kind of excuse that is frequently tendered is the excuse of ignorance. To accept this excuse in some circumstances but dismiss it in others, simply on the basis that, in the latter circumstances, the person in question should have known what he did not know, is bizarre; for this neglects the possibility that one has an excuse for not doing (or knowing) what one should have done (or known), the very possibility that gives rise to the question of how one can be to blame for ignorant wrongdoing in the first place!" (Zimmerman 2009, 254; see also Sher 2009, 20 & 81–82 & 85–86.)

Additionally, Zimmerman also mentions a fourth thesis, *attributionism* (i.e., answerability), which opposes the searchlight view and that Sher also mentions as one version of the minimalist position concerning the boundaries of the responsible self (see sect. 4.1.3; see also Sher 2009, 12–16 & 120–121 & 124–128). Whereas Sher rejected this thesis (in its basic form), Zimmerman suspects he may agree with it, given certain conditions. As also presented in section 4.1.3, attributionism is a view advocated by philosophers Thomas Scanlon and Angela Smith, among others, and one that Zimmerman describes to see that:

“[W]e can be responsible not just for acts but also for attitudes such as beliefs, desires, and emotions, even if we are unaware of the wrongness or foolishness of these acts and attitudes, as long as they reflect our rational judgments in such a way that it is appropriate, at least in principle, to ask us to defend them.” (Zimmerman 2009, 252.)

In other words, according to attributionism, it is the ability to make rational judgments that constitutes someone as a responsible person, and it is the particular contents of those individual judgments that make a person responsible (see also sect. 5.2; A. M. Smith 2010). Sher rejects this because, like the other theses, it is unable to accommodate the full range of what he considers our everyday judgments of responsibility (Sher 2009, 12–16; see also Zimmerman 2009, 252). For example, in the case of Alessandra leaving her dog in the van (Sher’s example case *Hot Dog*), an attributionist may say that Alessandra dealing with the problems at her children’s school reflects her judgment that she *has* good reason to do so, but this enables us to not consider what seems to make her behavior wrong – namely, leaving the dog in the van (Zimmerman 2009, 252).

Zimmerman, on the other hand, suspects that he may agree with attributionism, once some terminological differences are noted and limitations of attributionism are recognized (Zimmerman 2009, 258–260). I will explain what this means more specifically, but first I’ll introduce the remaining Thesis B – i.e., the qualified searchlight view – where Zimmerman and Sher most strongly disagree.

### 5.1.2 Failure of common intuitive judgments: The origination thesis

Thesis B, a *qualified version* of the searchlight view, is what Zimmerman himself accepts and Sher rejects. The reason why Sher rejects the view seems to be, ultimately, because it doesn’t fit with his intuitions about the nine example cases. For example, when briefly addressing specifically the Thesis B kind of searchlight view, Sher mentions that in the example cases, the agent’s lack of awareness that s/he is acting wrongly “simply does not appear to be traceable to any prior wrongful act or

omission”, and furthermore, even where there has been previous wrongdoing, an agent’s responsibility does not appear to depend on his or her previously having been *aware* that they were acting wrongly.<sup>88</sup> (Sher 2009, 33–39; see also Talbert 2011, 143.)

The reason why Zimmerman rejects Sher’s account of the epistemic condition of responsibility and is in favor of Thesis B – despite it not accommodating what they both consider the full range of our everyday judgments of responsibility – is that he is persuaded by an argument that it is true. Furthermore, he is disappointed in Sher for claiming that advocates of the searchlight view have paid little direct attention to defending it and which prompted Sher to engage in his “imaginative reconstruction” (as presented in sect. 3.4). Zimmerman emphasizes that he himself has provided an explicit argument for the view, and, even though Sher is apparently acquainted with it as he cites Zimmerman’s article where the argument first appeared, nevertheless he fails to properly engage with it. (Zimmerman 2009, 255–256.)

The argument that has convinced Zimmerman of the trueness of Thesis B – i.e., the qualified searchlight view – is rather complex and long but has been presented in a concise formal structure by him (1997, 2009, 2014).<sup>89</sup> Below is his central argument, in which the situation discussed is reminiscent of Alessandra forgetting Sheba, but where, as Zimmerman’s example, T = your act of turning on your stove, and D = your neighbor’s consequent death, where T somehow resulted into D (quoted directly from Zimmerman 2014, 1–2; see also Zimmerman 1997, 414–415; 2009, 256–257; Rudy-Hiller 2018, ch.2; Wieland 2017, 12):

- (1) T was wrong, but in doing it you acted from ignorance – that is, from lack of awareness – of this fact.<sup>90</sup>
- (2) One is culpable for acting from lack of awareness of some fact, and thereby for any consequences of so acting, only if one is culpable for one’s lack of awareness.

Hence

- (3) You are culpable for T, and thereby D, only if you are culpable for the lack of awareness, L, from which you did T.
- (4) One is culpable for something only if one was in control of that thing.

Hence

- (5) You are culpable for T, and thereby D, only if you were in control of L.
- (6) One is never directly in control of becoming aware of something of which one is currently unaware; that is, any control that one has over such awareness is only ever indirect.

- (7) If one is culpable for something over which one had only indirect control, then one's culpability for it is itself only indirect.
- (8) One is indirectly culpable for something only if that thing was a consequence of something else for which one is directly culpable.

Hence

- (9) You are culpable for T, and thereby D, only if there was something else (call it X) for which you are directly culpable and of which L was a consequence.
- (10) Whatever X was, it cannot itself have been an instance of ignorant behavior, since otherwise the foregoing argument would apply all over again to it; that is, whatever X was, it must have been some item of behavior of whose wrongness you were aware of at the time you engaged in it.

Hence

- (11) You are culpable for T, and thereby D, only if there was some item of behavior, X, for which you are directly culpable, of whose wrongness you were aware at the time you engaged in it, and of which T and D were consequences.

This argument is what leads to what Zimmerman calls the *Origination Thesis*:

“Every chain of culpability is such that at its origin lies an item of behavior for which the agent is directly culpable and which the agent believed, at the time at which the behavior occurs, to be overall morally wrong.” (Zimmerman 2009, 257; see also Zimmerman 1997, 423; 2014, 2.)

As far as Zimmerman can tell: ignorant behavior is rarely to be traced to a non-ignorant origin – and thus, the Origination Thesis implies that culpability is rare when an act is due to ignorant behavior. Thus, it seems that even the qualified version of the searchlight view – Thesis B – substantially focuses on the clause 2a, namely that a morally responsible agent is primarily one who is aware of the act's wrongness or foolishness.<sup>91</sup> In terms of our common practices that Zimmerman sees to match well with Sher's intuitions, he thinks they are too expansive: we are often too quick to blame people for their ignorant behavior.<sup>92</sup> (Zimmerman 2009, 257.)

### 5.1.3 Different kinds of agent evaluability?

Finally, briefly returning to the point of why Zimmerman suspects that he may agree with attributionism, once some terminological differences are noted and the limitations of attributionism are recognized. This is because Zimmerman suspects that he and Scanlon – who Zimmerman

explicitly refers to, but this seems to apply to Smith as well – may only be talking about possible different *kinds* or *modes* of moral responsibility, and may thus actually agree with each other. (Zimmerman 2009, 258–260; see also Zimmerman 2014, 3–5.)

The confusion rests on attributionists’ challenge to Zimmerman’s premise (4), where Zimmerman takes *freedom of will* or *control* as a necessary condition for culpability.<sup>93</sup> “Culpability”, for Zimmerman, is the negative mode of what he calls “appraisability”, and “laudability” is its positive mode. Instead of “appraisability”, Scanlon talks about “attributability”. According to attributionism, the particular contents of the individual judgments that makes a person responsible does not necessarily require any kind of freedom of will, as it suffices that the attitudes or traits – such as beliefs, desires, and emotions – that constitute a responsible person are merely expressed whether or not we are aware of those expressions and whether or not we are in control of them. Even though free will is very commonly viewed as a requirement for moral responsibility, there are some, like Scanlon, who challenge this. To use philosopher Robert Adams’s example that Zimmerman mentions: a graduate of Hitler *Jugend* is to be blamed for his vile beliefs, no matter how he acquired them (Adams 1985, 19). (ibid.; see also sect. 2.2.)

Zimmerman is willing to accept this thesis insofar as its limitations are recognized; namely, that a socially forced or indoctrinated moral inability is recognized to be crucially different from moral inability as a consequence of free deliberation – and which distinction reveals two different kinds of culpability. Zimmerman believes that a reaction of *punishment* is warranted and deserved only when the agent has freely and consciously done wrong. Zimmerman’s “appraisability” and Scanlon’s “attributability” may thus refer to different kinds of agent *evaluability*. If Scanlon agrees, then Zimmerman and he (and other attributionists, like Smith) may not be at odds. (ibid.; later on, in the discussion, Sher has also been explicitly classified as an attributionist, while Scanlon and Smith have been classified to represent the distinct category of answerability theorists; see sect. 2.2 & 5.4.)

In terms of challenges to his argument by Sher, Zimmerman mentions that nothing in Sher’s book indicates exactly which premise or premises of his argument Sher would necessarily reject (Zimmerman 2009, 257–258). It seems to me that Sher ultimately rejects the qualified version of the searchlight view because he thinks the overall conversation and our thesis about the epistemic condition should start with what our intuitions seem to say, whereas Zimmerman thinks the conversation and our thesis about the epistemic condition should start with his argument in support of the Origination Thesis. However, having made the possible distinction between the two kinds of moral evaluability between the qualified searchlight view and attributionism, Zimmerman suspects that this sort of talking past one another may be a wider problem within the field, without us properly

realizing it. In that vein, Zimmerman is willing to yield a compromise to Sher – one that is essentially the same he suspects Scanlon might accept: you can have your form of responsibility (without awareness) so long as it doesn't correlate with the agent's deserving the particular reaction of punishment (Zimmerman 2009, 260–261). As Sher says that the agents in the example cases “would definitely be blamed and might well be liable to punishment”, Zimmerman agrees that this may well be a common reaction the agents would face, and even yields that they might *deserve* some form of blame but remains firm that they would not deserve *punishment* (Zimmerman 2009, 260; Sher 2009, 24).<sup>94</sup>

Sher, however, is likely not satisfied with this proposition of accepting responsibility as a largely fragmented concept, as it would undermine his all-encompassing project to explain *all* of moral *and* prudential responsibility (Zimmerman 2009, 260–261).

#### 5.1.4 Conclusion

Preliminarily, I am myself prone to gravitate towards Zimmerman's views on the epistemic condition of moral responsibility. However, I suspect I would not fully agree with his thoughts on punishment being warranted and deserved only when the agent has freely and consciously done wrong. The degree to which I would agree: I do not think punishment is *deserved* without the agent having acted freely and consciously (if even then, given that determinism is true); but maybe blame is *warranted*, and possibly punishment, insofar as there are clear and well-proportioned pragmatic benefits that would have been well demonstrated with the form of blame or the form of punishment that is being suggested. However, I am not sure whether there actually are any kind of benefits to any form of blame or punishment for any *unconscious transgression*, be them benefits for the agents involved or the society, but it is a possibility.

My initial gravitating towards Zimmerman's views as well as the minor reservation are closely linked with what I present in chapter 6. Specifically, it is linked with my reservations about Sher's (and, in part, Zimmerman's) intuitions about the nine example cases (sect. 6.1), and the pragmatic point that emphasis *should* be put on us to building a community of epistemically informing each other in an open and compassionately non-judgmental deliberative space, rather than building a community of us negatively morally judging each other (sect. 6.3 & 6.4 & 6.6). It seems that the more we contribute towards the latter, the less likely we are to manifest the former.

## 5.2 Angela Smith's Attributionism

When considering the second minimalist position, *attributionism*, concerning the boundaries of the responsible self (in sect. 4.1.3), Sher connected the view particularly to the views of philosophers Thomas Scanlon and Angela Smith. As Smith has written an explicit review of Sher's book, wherein she also reflects and outlines her own views – and as also Zimmerman mentioned that his view may be compatible with attributionism if attribution and appraisal are properly distinguished – it seems especially fitting to examine what she thinks about Sher's arguments.

Smith sees the great advantage of Sher's account being that it neatly connects the nine example agents to their wrongdoings via their constitutive features, and she is especially pleased with Sher's overall critique of the searchlight view. Still, she does mention a minor quibble that Sher seems a bit hasty to dismiss the claim of it being unfair to hold an agent to a moral demand they had no way of complying with (see 3.4.2.2–3.4.2.3). This makes sense, as Smith's own view would like to figure out if an agent's actions represented a (normative) evaluative judgment on their part, to evaluate their responsibility. If they do reflect the agents evaluate judgment – whether or not the agent was aware of those judgments or in control of the expression of those judgments – and the act itself is normatively immoral, then the agent would be responsible and those judgments would define the boundaries of the responsible self.<sup>95</sup> Smith's primary concern, however, has to do with the "causal nature" of Sher's account. Further, there are classificatory differences between the two. (A. M. Smith 2010, 516–522.)

Sher classified Smith's view as a minimalist position (see sect. 4.1.3.1 & 4.1.3.4). While Smith is glad that Sher abandons the maximalist position (4.1.3.1), as well as the first minimalist position of volitionism (which, Smith mentions is embodied in the searchlight view) (4.1.3.3), she is perplexed that he would call her position minimalist. This is due to Smith's view encompassing not only an agent's reason-responsiveness or evaluative judgments, but also any conscious or unconscious desires, beliefs, attitudes, feelings, awareness, and unreflective patterns of thought insofar as these are expressions of the agent's evaluative judgments (cf. 4.1.3.4). Furthermore, Smith is perplexed that Sher would classify Harry Frankfurt's hierarchical view as well as John Martin Fischer and Mark Ravizza's view as intermediate positions (cf. 4.1.3.5). Smith sees both positions closer to the minimalist position of volitionism than her own account. (A. M. Smith 2010, 521–522.)

Regardless of the classificatory differences between Sher and Smith, Smith's concern focuses on how Sher argues that her position is not expansive enough. Sher thinks that in addition to noting the agent's reason-responsiveness we need to also account for the causal structure that sustains it

(along with the causal structure sustaining the agent's consciousness). He reasons that in many cases of unwitting wrongdoers, there does not seem to be any evaluative action at play on the part of the agent, which prevents us from making sense of the (intuitive) responsibility of the agent. For example, as Alessandra *forgot* about Sheba, and thus it does not seem to represent her evaluative judgments about Sheba, it seems she would be pardoned, unlike what Sher's intuitions would suggest. However, Smith thinks Alessandra may still be regarded responsible under her account, given appropriate background assumptions – and if given different background assumptions, she thinks Alessandra should not be held responsible. (A. M. Smith 2010, 522–523; see also sect. 4.1.3.4; Talbert 2011, 143–144.)

Smith sees that an agent's tendency to notice and remember things can be signs of what they judge to be important or significant. Thus, Alessandra forgetting Sheba may be a sign of an evaluative judgment on Alessandra's part, and that she could thus be responsible. However, Sher's objection is that Alessandra is distracted by the urgency and the stressful nature of the situation, which, due to her constitutive features, forces her to forget Sheba – and hence her forgetfulness does not represent an evaluative judgment (see 4.1.3.4). Given this, Smith does not see why we should hold Alessandra responsible. Referring to Sher's analogy of malfunctioning constitutive sparkplugs of cars and constitutive computer parts being akin to the constitutive features of humans (see 4.1.3.2), Smith is perplexed why we should hold Alessandra responsible for a mere glitch in her psychological system. Similarly, why should we blame the whole car or our computer for some uncharacteristic glitch in their parts? (A. M. Smith 2010, 523; see also Talbert 2011, 149–150.)

The basic problem Smith describes of Sher's account is that a causal connection between our awareness and a glitch in our psychological system does not establish the right *kind* of connection between an agent and their wrongdoing to justify us regarding the agent responsible. Smith suspects that if we were to trace an agent's failure strictly to their physical structures and mechanisms (e.g., to their brain), we would not hold the agent responsible. She suspects that *this* would be the majority intuition, because physical structures do not reflect anything about us as a valuing agent. But insofar as the act (and the physical structures behind the act) represent the agent's evaluative judgments, and hence s/he could in principle be asked to defend those judgments, the agent can be regarded responsible. Hence, the connection between an agent and their wrongdoing, in Smith's account, cannot be merely causal but it also needs to be *rational* – and Sher's account is unable to make this distinction. (A. M. Smith 2010, 524.)

Still, Smith concludes that the dispute does not seem to be serious, but merely a family dispute. She sees Sher's overall project to be commendable: there needs to be some connection between an

agent's wrongdoing and their constitutive attitudes, dispositions, and traits. Alas, in her view, that connection cannot be merely causal but also needs to be rational.<sup>96</sup> (A. M. Smith 2010, 524; for detailed accounts of Smith's position, see A. M. Smith 2004, 2005, 2008, 2012.)

Much like in the case of Zimmerman's qualified searchlight view, I find Smith's attributionism (i.e., answerability) also much more promising than Sher's FEC. However, the one reservation I have is likewise similar, and foreshadowing my examination in chapter 6: I think we should hold an agent responsible for their evaluative judgments – whether conscious or unconscious, or in their control or not – only insofar as we have reason to believe that there are well proportioned pragmatic benefits for doing so, whether for the agent or the society (or, in minimum, that there are no serious disadvantages for doing so). At the same time, however, regardless of whether there are such benefits, it may be that Smith's view is also closer to (some) majority intuitions than Sher or Zimmerman. Notably, it seems Zimmerman's views along with those presented in chapter 6 are more so normative (or at least have normative revisionist implications), whereas Sher and Smith are aiming to be more so descriptive (see also sect. 5.4).

### 5.3 Remarks in Reviews: Psychological Limits, Doubts, and Controversies about Control

Sher's book has been reviewed in a number of journals by various philosophers. Here, I introduce some remarks that I have come across in some of the reviews and that I think are noteworthy – in addition to the previous reviews by Michael J. Zimmerman and Angela Smith.

In his review, philosopher Bruce N. Waller asks that “even if we accept the claim of moral responsibility in searchlight cases, does it follow that because the same mechanism is involved in both there is moral responsibility in the more problematic cases?” Sher's expansion of the searchlight view via PEC and ultimately FEC requires that we accept that there is some connection between claims of moral responsibility in searchlight cases and the nine example cases that he provides to illustrate why non-conscious aspects would also matter. However, this doesn't seem to be the case, necessarily, and Waller presents a scenario as a thought experiment that would seem to indicate this. (Waller 2014, 642.) Waller's scenario tries to illustrate why Sher's intuitions in his example cases may be too quick to judge:

“Imagine a runner who has just completed three eight-hundred meter sprints in quick succession, and now must run another. She runs the final sprint at a pace that is severely substandard compared to her usual swift pace, and we blame her for her dilatory effort. ‘That’s not fair’, she will reply. ‘I ran as fast as I could. It’s not fair to blame me for running less successfully when I’m exhausted from three hard sprints’. ‘You are still to blame’, we insist, ‘after all, the constitutive elements that were involved in your swifter first sprint are exactly the same elements involved in your last subpar effort’. ‘Of course they are’, she retorts, ‘I didn’t imagine that my last sprint was subpar because a demon had taken possession of the constitutive factors that enable me to run. It was subpar because those “constitutive elements” were already run down. Of course the subpar effort was my own; but that doesn’t mean I deserve blame for it.’ This is precisely the situation some of Sher’s cases involve.”<sup>97</sup> (Waller 2014, 642–643.)

Via this example, Waller notes how similar problem may apply to several or all of Sher’s example cases. He discusses, as an example, Sher’s example case *Hot Dog*; the case where Alessandra drives to pick up her children at their elementary school, gets distracted by the school administrators’ bungling, and forgets the family dog Sheba in the hot van for several hours. Waller notes that settling the school administrators’ bungling requires substantial investments of rigorous deliberation on Alessandra’s part, and this makes her forget Sheba. Alessandra does not consciously recall Sheba being in the van, and, Waller notes, “of course the same constitutive elements that enable Alessandra to deliberate consciously (and usually very well) are in this instance the source of her deliberative failure (her failure to bring the important factor of Sheba into her conscious deliberations).” Thus, in Waller’s view, even though Sher may be right that *intuitively* we may feel that Alessandra deserves blame for forgetting Sheba, perhaps our intuitions shouldn’t be so quick to judge here. Sometimes our constitutive deliberative elements (of which we are largely unaware of) may, as a consequence of a particular situation, be compromised in an equivalent way to the ability of a runner to run three consecutive eight-hundred-meter sprints in the same quick pace. It becomes physically and psychologically impossible. (Waller 2014, 642–643.)

Waller notes that Alessandra may be in a state of what psychologist Roy Baumeister calls ‘*ego depletion*’: a state where her capacity for deliberation is severely limited due to her being mentally encumbered in the situation.<sup>98</sup> By making this overall notation, Waller is saying that the *same* constitutive deliberative faculties of an agent may function well in one situation, when the agent is well rested and in good health, but may function poorly in another, when the agent is encumbered or exhausted. As a result, it may not be so clear as Sher thinks that his example case(s) would require

responsibility – and hence blameworthiness – outside the confounds of some version of the searchlight view, when we take note of these kinds of cognitive boundaries in human deliberation, and consequently judge them more charitably.<sup>99</sup> (Waller 2014, 643.)

A somewhat similar sentiment of doubt about the limits of the searchlight view is also briefly echoed by philosopher Dana Kay Nelkin, stating as an example of Sher's example case *Jackknife* that "depending on whether Father Poteet's case is filled out with details about what was likely to have happened if he had stopped abruptly instead, we might be tempted to absolve him of responsibility altogether" (2011b, 676 & 678; cf. sect. 3.3.2).<sup>100</sup> Likewise, philosopher Matthew Talbert echoes this sentiment, illustrating the point via Alessandra: we do not know anything about her background nor about the specifics of the situation, e.g. what the hassle at the school was about, which may contain further details that would affect our intuitions (2011, 147–148).<sup>101</sup> Talbert also notes that as Sher doesn't distinguish between a "moral wrongdoing" concerning a mental lapse from a wrongdoing concerning a deliberate negligence, it is not very clear how strong a wrongdoing the former would be in comparison (2011, 150–151). It can even seem tragic rather than blameworthy (Talbert 2011, 150–151). Thus, overall, Sher's intuitions about these cases do not necessarily enjoy such an infallible status as Sher seems to take them to enjoy. Even if we accept the agents in the example cases to be intuitively responsible, some version of the searchlight view may still suffice to explain those intuitions, after we specify further details – or background assumption – about the cases.

In Sher's argument about the applicable standard in clause 2a of FEC – in which he endorses evaluating the agent against a similar situation via his own variant of the reasonable person standard (see sect. 4.1.2) – he encourages us to focus on the agent's constitutive attitudes, dispositions, and traits, distinguished from the agent's situation. Applying some of Nelkin's thoughts to my own: the examination of where to draw the line between the agent and their situation may run into some trouble if we were to examine some relevant psychology. The trouble being that it may be very hard to define the constitutive properties of a person, psychologically, even if we find some demarcation intuitively appealing. In Sher's account, there is no obvious reason why some aspects of the agent's psychology would not be considered constitutive. Thus, Sher's case further rests on arguments of what exactly should be considered constitutive of agent's psychology and why. Unless further extrapolation is made, the constitutive features might be so expansive as to practically include the whole of agent's psychological make-up. This might allow features such as *moral blindness* – and similar properties – to be counted as constitutive of an agent. At which point maybe the searchlight view (i.e., Sher's clause 1), or as Nelkin says "what can fairly be expected of the person", is all that matters. Sher even

himself notes that there is a lot of disagreement about where the line should be drawn, and is himself rather vague about the demarcation, but remains committed in his approach (see sect. 4.1.3.2). Specifically, there is a lot of disagreement about the matter in the legal context, let alone in the moral context, into which Sher is bringing concepts from the legal context (most relevantly “similar situation” and “the reasonable person standard”). (see Nelkin 2011b, 677–679.)

Furthermore, Nelkin notes that Sher’s applicable standard states that moral demands are only directed at agents in their capacity as reason-responders (see sect. 4.1.2.2). The relevant cognitive capacities Sher takes to include, for example, the agent’s disposition to notice various features of their surroundings (ibid.). This, Nelkin notes, seems to be in tension with Sher’s demarcation of the boundary between the agent and their situation. For example, if we consider an agent’s habitual fear *to limit* their cognitive capacities, we cannot also consider it to be *part of* their constitutive features but rather a part of their situation. Thus, for example, it could be interpreted that the agents in the nine example cases are not responsible, insofar as their failure was due to their disposition to not notice relevant features of their surroundings (or if they fulfil the other criteria outlined at the end of section 4.1.2.2). (Nelkin 2011b, 679.)

However, having said everything in the last two paragraphs, we may charitably interpret Sher’s list of relevant cognitive capacities to refer to agents who are *significantly* limited in their abilities. Thus, usual fear would not be included in the list, but, for example, features such as cognitive disabilities, clinical memory loss, and severe psychological illnesses would. All cognitive features not included in severe limitations would be considered constitutive. Of course, a grey area would still remain, but this interpretation would at least enable some clear distinctions. (see sect. 4.1.2.2; see also 4.1.3.2; Talbert 2011, 146–147.)

As a final critical point of Sher’s account, Nelkin (2011, 679–680) asserts that Sher is committed to a *very controversial* view of the control condition, in other words the “ought implies can” principle (see sect. 3.4.2 & 4.3). As Nelkin (2011, 679–680) summarizes, the principle states that one is obligated to do something only if one has control over it. Sher’s view goes directly against this as, in his view, one can have moral obligations to exercise certain psychological capacities even when those exercises are not consciously chosen (i.e., in a sense, even when one has no control over them). Despite Sher’s attempt to justify abandoning the condition, the topic is left quite open (see sect. 4.3). If Sher’s account requires us to abandon the control condition, and the control condition is closely connected to the searchlight view, Nelkin takes this to mean that the searchlight view or some variation of it might come on top after all.

Sher himself answers a question of might there be a double standard when he trusts his intuitions so strongly in formulating his case, but at the same time dismisses the strong intuitive appeal of the control condition: he claims the difference lies in that the example cases he uses to justify our (or his) intuitions about the role of non-conscious factors in responsibility are relatively uncontaminated by theory, whereas the control condition is deeply bound up with theoretical commitments that should be abandoned (Sher 2009, 152; see also sect. 4.3).

## 5.4 Outlines of Alternative Accounts of the Epistemic Condition

Since Sher's 2009 account, a prolific discussion about the epistemic condition has emerged. As the discussion includes various kinds of considerations about the condition from many different angles, it is, of course, impossible to thoroughly summarize it all here. However, in this section, I provide a very brief outline of where the discussion seems to be at the moment. Three sources in particular seem promising, as they all present their own summaries of the discussion: Rudy-Hiller 2018; Talbert 2016, ch. 5; and Wieland 2017. The following can be viewed as a continuation of section 2.2, adding to the recent history of discussion about moral responsibility, this time specifically after the focus on the epistemic condition has emerged.

The two main positions regarding moral responsibility more generally were introduced in section 2.2: accountability and attributability. According to Wieland (2017, 5 & 5n7), most philosophers in the debate about the epistemic condition are concerned with responsibility as accountability as opposed to attributability. Hence, it seems safe to assume that most of the views are merit-based (i.e., desert-based, backward-looking views). Thus, the central question in the current field seems to be: *when* – in a normative sense – is it reasonable to expect agents to do (or have done) better, so we may know when they should be excused from responsibility?<sup>102</sup> As we have already seen, there is substantial disagreement on the details of the epistemic condition.

As presented in chapters 3 and 4, Sher thinks that there can be 'responsibility without awareness'. However, the more usual position appears to be that *some* awareness is required. I outline the related discussion in section 5.4.1. Relating to different positions about the role of awareness in responsibility, we can then distinguish five main types of views about the epistemic condition. Broadly speaking, we can distinguish the *revisionist accounts* of the epistemic condition (e.g., Zimmerman's being a prominent account) from four types of accounts that aim to answer the revisionist accounts (e.g., Sher's being one example, and Angela Smith's another, with two further

types out there). After introducing some useful terminology to approach them with (5.4.2), I outline the revisionist account (5.4.3) and the four types of answers to it (5.4.4). In section 5.4.5, I provide a guideline to how to approach chapters 6 and 7, where I present my own answer to Sher, representing a further view of the epistemic condition.

In the discussion it seems that positive and negative moral responsibility are usually viewed symmetrically – contrary to, for example, Sher who sees them to be asymmetrical (see sect. 4.2) – but, once again, I focus on addressing the negative side much like the discussion in the field in general.

#### 5.4.1 The role of awareness

Unlike Sher, it seems that most philosophers do think some kind of awareness is required for responsibility for ignorance. However, the question about the necessity of awareness is only one of three issues that are debated. The other two take awareness to be required, but then debate on what the *contents* of the required awareness would be (i.e., what the target agent needs to be aware of) and on what *kind* of awareness would be required (i.e., what mental states are involved and in what way they need to be entertained). (Rudy-Hiller 2018, ch. 1.) These two lines of inquiry concerning awareness – contents and kinds – are outlined in sections 5.4.1.1 and 5.4.1.2, respectively.

##### 5.4.1.1 Contents of awareness

Concerning the contents of awareness, there are four areas of awareness that have been discussed: awareness of action, moral significance, consequences, and alternatives. Each of these, or some combination of these, may be viewed as either necessary or sufficient conditions for the epistemic condition, depending on the specific theory or case. Or, it may be viewed that none of them are necessary. (Rudy-Hiller 2018, sect. 1.1.)

When someone is *aware of action* (or omitting an action) they are aware of performing it under appropriate description. For example, if we view that Alessandra needs to be aware of her action to be responsible for Sheba languishing in the car, Alessandra needs to be aware of her omission to tend to Sheba at the time that omission occurs. Or, for example, amerika – who robs the bank in Sher's example case *Bad Weather* – needs to be aware of him robbing the bank when it occurs. (ibid.; sect. 3.3.2.)

*Awareness of moral significance* is in the summarizing literature distinguished to refer to two types of awareness: *de dicto* awareness and *de re* awareness. *De dicto* awareness refers to an agent

having a belief about the action being all-things-considered wrong (i.e., concerning what an agent thinks is morally relevant to their action). *De re* awareness refers to an agent having a belief about the presence of whatever features *in fact* make the action wrong (i.e., the action's wrong-making features), regardless of whether the agent is *also* aware of the moral significance of those features. For example, for Alessandra to have been *de dicto* aware, she would have likely been aware that it is wrong to allow a dog to suffer (in a hot van for hours). And, for Alessandra to have been *de re* aware, she would have been aware that Sheba is to be let out of a hot van (and she may or may not have been aware that this is because it is wrong to allow a dog to suffer in a hot van for hours). A position that considers *de dicto* awareness necessary for the epistemic condition entails that moral knowledge (or at least moral belief) is necessary for responsibility (Sliwa 2017), whereas a position that considers *de re* awareness necessary denies the necessity of moral knowledge (Harman 2011, 2015; Talbert 2013). The question of whether moral knowledge is required, and thus what kind of awareness of moral significance is required, is a key dispute concerning the epistemic requirements on blameworthiness. (Rudy-Hiller 2018, sect. 1.1; Wieland 2017, 18–19.)

*Awareness of consequences* refers to awareness or belief of an event occurring as a result of an action. For example, in the case of Alessandra this awareness or belief might refer to awareness that leaving Sheba in the van would (or might) lead to the consequence of the dog languishing. There is disagreement on whether awareness of consequences need to be specific (i.e., that an agent has to believe an event exactly the kind that occurs to possibly result from their action; “actual foresight”) (Zimmerman 1986; Vargas 2005), or if a more general content of consequences suffices (i.e., “reasonable foreseeability”) (J. M. Fischer & Tognazzini 2009; King 2017, 272; Nelkin & Rickless 2017a, 126). For example, according to the former, Alessandra would need to believe Sheba might *languish* as a consequence of her leaving her to the car; whereas according to the latter, Alessandra would need to believe Sheba might be *harmed in some way* if leaved in the car. This debate relates to a broader dispute of whether the epistemic condition requires occurrent awareness (see sect. 5.4.1.2 & 5.4.3) or, rather, the capacity to gain the requisite awareness (sect. 5.4.4.2 & 5.4.4.4). (ibid.)

*Awareness of alternatives* refers to awareness of alternative actions that could be taken in a given situation. For example, if Alessandra was not aware that there were other (morally permissible) options available to her than to leave Sheba in the van, she would not be responsible (Levy 2011, 111). (ibid.)

Overall, whenever we are talking about awareness, we may be referring to one or more of the abovementioned four types of contents of awareness. There appears to be more or less dispute about all of the four – about awareness of action, moral significance, consequences, and alternatives – in

terms of what (combination) the epistemic condition would require. Lastly, there is the position that, for example, Sher (2009) holds, that no awareness is necessarily required (see ch. 3 & 4).

#### 5.4.1.2 *Kinds of awareness*

Concerning kinds of awareness, there are two questions that are discussed: firstly, *what* mental states the target agent needs to entertain in order to possess the relevant awareness, and secondly, *how* must those mental states be entertained. (Rudy-Hiller 2018, sect. 1.2.)

The candidates for the mental states that the target agent needs to entertain, for them to possess the requisite awareness are – in the order of demandingness – knowledge, or at least reasonable or justified belief, or at least (coincidentally) true belief, or at the very least some type of belief.<sup>103</sup> That is, for example, for Alessandra to satisfy the epistemic condition, she needs to have *known* (or, depending on the theory, have at least held some kind of belief of) what she was doing (awareness of action) and, potentially, she needs to have known (or had some kind of belief) about her actions moral significance, potential consequences, and possibly about other permissible alternative actions she could have taken. (ibid.)

Regarding how these mental states and their contents in question need to be entertained, there is dispute between two views: (1) they need to be entertained *occurrently*, or (2) it suffices they are entertained *dispositionally*. (ibid.)

On the one hand, *occurrentists* argue that an agent satisfies the epistemic condition only when, at the time of action, they consciously believe that their action is wrong and consciously contemplates some of its consequences (e.g., Zimmerman 1997, 421–422). Otherwise, an agent is ignorant of relevant considerations, and so has an excuse for their wrongdoing. This view plays a central role in, for example, Zimmerman's views; or what are now known as *revisionist* arguments against commonsense attributions of responsibility (sect. 5.4.3). (ibid.)

On the other hand, *dispositionalists* argue that the occurrentist interpretation of the epistemic condition is too strict and would let too many intuitively blameworthy wrongdoers off the hook. Rather, they think that tacit, dormant, dispositional, or unconscious beliefs can, at least in many cases, amount to the kind of awareness required for moral responsibility. This view seems to be especially present in weakened internalist positions (sect. 5.4.4.1), but in different ways also in others (sect. 5.4.4). (ibid.; see also Rudy-Hiller 2018, n3.)

### 5.4.2 The orthodoxy

Before outlining revisionism and the four types of views against it, Wieland (2017) introduces some useful terminology to approach them with. Further, he argues that some main figures in the debate accept what he calls the ‘*Orthodoxy*’, comprising of core agreement about the full set of claims presented in this section that can function as a useful mirror for the broader discussion (Wieland 2017, 9).

Philosopher Holly Smith (1983) has made a useful distinction between ‘benighting acts’ and ‘unwitting acts’. Benighting acts refer to omissions to inform oneself, while unwitting acts refer to subsequent ignorant behavior that follows from benighting acts. The former can either induce or perpetuate ignorance about the permissibility of unwitting acts. Following Wieland (2017, 6), these may be referred to in their temporal sequence as follows:

A1: benighting act

A2: unwitting act

Firstly, according to Wieland (2017, 9–12), there is agreement that excuse by ignorance can render S (i.e., a moral agent, a subject) blameless. Secondly, there is agreement that in cases where S is not ignorant that his/her action is wrong, excuse by ignorance is (obviously) not on the table. Thirdly, there is usually agreement that in cases where S is ignorant that A2 is wrong (either because of factual or moral ignorance), it is accepted that:

- (i) S is blameworthy for A2 only if S is blameworthy for her ignorance that A2 is wrong;  
and
- (ii) S is blameworthy for her ignorance that A2 is wrong if and only if S is blameworthy for a benighting act A1 (at least one such act) that led to A2.

Cases where (i) and (ii) apply are cases of *derivative* (or *indirect*) blameworthiness: blameworthiness for A2 derives from blameworthiness for A1. Although there tends to be agreement of these, conditions for the latter are debated. Additionally, there can be cases of *original* (or *direct*) blameworthiness: blameworthiness for A2 does not derive from blameworthiness for A1. Agents who are not ignorant of the wrongness of their conduct, but do it nonetheless, may be directly blameworthy. (Wieland 2017, 9–10.)

Following Wieland (2017, 10) and the sequence of A1–A2, we may refer to the possible corresponding responsibility as follows:

B1: blameworthiness for benighting A1 (or the ensuing ignorance)

B2: blameworthiness for unwitting A2 (or a consequence of A2)

Utilizing these terms, Wieland (2017, 10–11) argues that many philosopher – though not all – who defend derivative accounts of responsibility accept that derivative blameworthiness involves the following five claims:

Claim 1: If B1, and S has no further excuses for wrongful A2, then B2.

Claim 2: B2 only if B1.

Claim 3: B2 (partly) because B1.

Claim 4: Claims 1–3 apply to factual and moral ignorance.

Claim 5: Claims 1–4 are to be understood in terms of accountability.

These five claims are instructive to approach also philosophers who defend non-derivative accounts of blameworthiness. However, conditions for B1 is where most of the debate has focused thus far, and it is indeed the focus in this outline as well. (Wieland 2017, 10–12.)

#### 5.4.3 Internalism, revisionism, and akrasia

*Internalism* (i.e., volitionism, closely related to the searchlight view and that Sher contrasted with attributionism; see sect. 4.1.3.1 & 4.1.3.3) views that conditions for B1 and B2 are exactly the same, namely that blameworthiness requires witting wrongdoing. Thus, the view entails that: S is blameworthy for benighting act A1 *only if S believes that A1 is wrong* (or s/he is blameworthy for his/her ignorance about this). *Revisionism* (of the Orthodoxy or our common intuitions about responsibility) is what can follow from internalism as argued by Zimmerman (1997, 2008). In this section, I present some further details on this position (the general outline of Zimmerman’s position having been presented in section 5.1.2), and the next section presents four main categories of response to it. (Wieland 2017, 12.)

As Zimmerman demonstrated, internalism is on the road to regress (see sect. 5.1.2). Wieland (2017, 12) has reconstructed this *regress argument* as follows:

- (1) S is blameworthy for A2 only if
- (2) S believes that A2 is wrong, or
- (3) S is blameworthy for her ignorance that A2 is wrong
- (3) only if

- (4) S is blameworthy for the past omission A1 that resulted in her ignorance that A2 is wrong
- (4) only if
- (5) S believes that A1 is wrong, or
- (6) S is blameworthy for her ignorance that A1 is wrong
- (6) only if etc.

The regress entails Zimmerman's origination thesis (sect. 5.1.2; Wieland 2017, 12):

"Every chain of culpability is such that at its origin lies an item of behavior for which the agent is directly culpable and which the agent believed, at the time at which the behavior occurs, to be overall morally wrong." (Zimmerman 2009, 257; see also Zimmerman 1997, 423; 2014, 2.)

In other words: blameworthiness requires *akrasia*, that is someone consciously acting against their better judgment, i.e. requires occurrent *de dicto* awareness of the wrongness of one's action (see sect. 5.4.1). Or: "S does X akratically iff, at the time of X, S believes that she should not do X, i.e. that X is all-things-considered wrong". Although it is somewhat debated whether *akrasia* is only necessary for blameworthiness or also sufficient – depending on whether control is also necessary – Zimmerman's regress argument rests on its necessity. (Rudy-Hiller 2018, ch. 2; Wieland 2017, 12–13.)

Further, Zimmerman argues that *akrasia* rarely occurs (2008; sect. 5.1.2). Arguably, we tend to always believe that our actions are all-things-considered permissible. This brings about the *revisionist implication*: clauses (2) and (5) are hardly ever satisfied. Similar arguments to Zimmerman's have been presented by, for example, Gideon Rosen (2004) and Neil Levy (2011). Rosen (2004) adds an epistemic twist: he does not say (2) and (5) are rarely satisfied, but that it is difficult, if not impossible, to determine that they are, and hence we hardly ever *know* whether agents are blameworthy for anything. (Wieland 2017, 13–14.)

Those adhering to the Orthodoxy accept the regress at least until clause (4), it being motivated by control considerations and the *falsity of doxastic voluntarism*: we lack direct control over our beliefs, that is, we can't decide at will what to believe (Wieland 2017, 12). At the same time, the Orthodoxy denies the revisionist implication and argues that agents are blameworthy in more cases than what Zimmerman and other internalists are leading us to believe.

#### 5.4.4 Four answers to revisionism

Broadly speaking, there can be distinguished four types of answers to the revisionist implication of the regress argument. These positions have been labelled a bit differently by Wieland (2017) and Rudy-Hiller (2018). Mentioning Wieland's categories first followed by Rudy-Hiller's, they correspond as follows: [I] internalist-friendly responses, i.e. weakened internalism; [II] externalist responses, i.e. accounts of ignorance and epistemic vices; [III] orthodoxy-breaking responses (internalists and externalists both accept the orthodoxy), i.e. capacitarianism; and [IV] quality-of-will accounts of blameworthiness for ignorance (as opposed to accounts of prior blameworthy conduct), i.e. quality-of-will accounts. Here, I will adopt the latter terms by Rudy-Hiller. Further, there could be distinguished accounts of *moral responsibility skepticism*, which are largely omitted from examination by both Wieland and Rudy-Hiller (cf. Caruso 2018), though revisionism may be viewed as a form of skepticism. These are omitted from this examination also, yet do briefly come up elsewhere in the thesis (e.g., sect. 6.1.2 & 6.6.3).

Below, in sections 5.4.4.1–5.4.4.4, I briefly summarize the four answers, respectively, in terms of how they approach the revisionist implications. Wieland's reconstructed regress argument presented in the previous section above should be a useful reference. I also list some philosophers that Wieland mentions to have advocated the positions, and provide my initial thoughts on the positions. These will follow and summarize Wieland's account (2017), but see Rudy-Hiller (2018) for an alternative, more in-depth exposition, and Talbert (2017, ch. 5) for a yet additional outline.

##### 5.4.4.1 Weakened internalism

These internalist-friendly responses accept (1)–(6) of Wieland's reconstruction of Zimmerman's regress argument, but propose a more liberal reading of (2) and (5), avoiding the revisionist conclusion in part. (Wieland 2017, 14.)

The idea is that blameworthiness can be traced back not only to *akrasia*, but also to other mental states. For example, even if one is unaware that their act is wrong when performing it, one might nonetheless suspect that it is, or hold the belief unconsciously that it is, or have sufficient (even though indecisive) evidence that it is. If this sort of view is accepted, then Zimmerman's criteria may be satisfied much more frequently. (Wieland 2017, 15.)

According to Wieland, this sort of view has been advocated by, for example, Haji 1997, Peels 2011, and Robichaud 2014. (ibid.)

Personally, I have doubts about this position because it appears that if we are epistemically honest, or properly skeptical, this may very often be our predicament when making choices. At least I personally have these kinds of suspicions all the time, concerning choices to all sorts of directions, with different seemingly valid arguments supporting different choices. The present discussion about what would be the right choice of view about the epistemic condition being only one example. The work required to make even a choice that *appears* right is thus arguably a great ordeal, the success of which to some significant degree rests on *luck*. Consequently, if weakened internalism was accepted, it would seem I could be held responsible very arbitrarily for some things yet not for some other things.

#### 5.4.4.2 *Ignorance and epistemic vices*

These externalist responses deny (5)–(6) as necessary for (4) in favor of an alternative sufficient condition for (4) relating to epistemic vices. (Wieland 2017, 14.)

The idea is that blameworthiness may trace not only to akrasia, but also to the exercise of epistemic vices (e.g., laziness, arrogance, incuriosity, dogmatism). And these vices can be sufficient for blameworthiness; i.e., even if the agent is not aware that they are doing anything wrong, they can be blameworthy if their ignorance is due to their vices. However, some externalists add that history of S's vices is also relevant. According to these views, S may be blameless for A1 if they had no opportunities to be less vicious (e.g., due to warped upbringing and non-virtuous surroundings). Thus, with the historical condition, the view is that S is blameworthy for A1 *if A1 is due to S's epistemic vices and S had normal opportunities to develop his/her virtues*. (Wieland 2017, 15–16.)

According to Wieland, this sort of view has been advocated by, for example, Montmarquet 1999, FitzPatrick 2008, 2017 (accepting the historical condition), and Talbert 2017 (without accepting the historical condition). (ibid.)

Personally, I view the variation with the historical condition more promising, but even then – foreshadowing chapter 6 – I have doubts about this position because it seems to ignore the pragmatic possibility that virtue might be better taught and spread without appealing to responsibility. Also, it is unclear how we would confirm someone having had or having been deprived of “normal” opportunities.

#### 5.4.4.3 *Capacitarianism*

These orthodox-breaking responses deny (2)–(3) as necessary for (1) in favor of an alternative sufficient condition for (1). (Wieland 2017, 15.)

The idea is that S's blameworthiness for A2 does not imply that S is blameworthy for her ignorance that A2 is wrong. Thus, S can be blameworthy for A2 without requiring any other blameworthiness fact to be true of S. Instead, it suffices for blameworthiness that S's unawareness of relevant aspects of their situation falls below a cognitive standard that applies to them, given their cognitive and volitional abilities and the situation they are in. (Wieland 2017, 16–17.)

As established, Sher (2009) represents this type of account (ch. 3 & 4). According to Wieland, another prominent advocate is Clarke 2014. (*ibid.*)

As chapter 6 especially underlines, these views appear to rely on suspect intuitions. Indeed, in addition to what I discuss there, and what has been discussed thus far, Neil Levy (2017) has provided an error theory for the intuitions that motivate Clarke (see also Wieland 2017, 17–18).

#### 5.4.4.4 *Quality of will*

Accounts focusing on quality of will (rather than on prior blameworthy conduct) deny (4) as necessary for (3) in favor of an alternative sufficient condition for (3) (and for a range of things). (Wieland 2017, 15.)

Different quality-of-will theorists differ in details, but they are united in their denial that blameworthiness for unwitting acts is derivative (i.e., in particular, they deny that blameworthiness for unwitting acts is to be explained by blameworthiness for benighting acts). Instead, blameworthiness for unwitting acts – and for many other things, e.g., benighting acts, failures to notice, remember, and other attitudes – is to be explained by a lack of moral concern. This is similar to the epistemic vices account except the focus is not on the lack in epistemic attitudes but moral. (Wieland 2017, 18–19; see also Wieland 2017, 20–22.)

According to Wieland, some variations of this account have been provided by, for example, A. M. Smith 2005, Arpaly 2003, and Björnsson 2017. The following approximations of these views may serve as examples of the variety in quality-of-will accounts: As established in section 5.2, Smith approximately views that S can be blameworthy for an attitude, including a failure to recognize or remember something, if S's attitude is due to a lack of moral concern on S's part. Arpaly approximately views that S is blameworthy for some unwitting A2 if, and only if, A2 is due to a lack

of concern for the features that make A2 wrong (i.e., lack of *de re* concern<sup>104</sup>; see sect. 5.4.1.1). And Björnsson approximately views that S is blameworthy for X when and insofar as X is explained by S's quality-of-will falling short of what can reasonably be expected. According to Wieland, many others have also defended variations of the quality-of-will approach, including many writers in Robichaud & Wieland 2017 (namely Talbert, Harman, Zimmerman, Mason and Wilson, Alvarez and Littlejohn). (Wieland 2017, 18–19.)

Personally, I view these positions comparatively promising out of the four answers to revisionism. However, like in the epistemic case (sect. 5.4.4.2), I am likewise uncertain whether appealing to responsibility is the best way to elicit or teach moral virtues. I am also cautiously sympathetic towards revisionism itself, as well as more largely towards moral responsibility skepticism. However, some of the quality-of-will accounts are interested in responsibility as attributability rather than accountability (Wieland 2017, 19n21), which may partly explain why I find them the most promising: they may be interpreted in terms of a forward-looking consequentialist view of responsibility, instead of a backward-looking merit-based view of responsibility (cf. sect. 2.2; see also Caruso 2018, ch. 1).

#### 5.4.5 Guidelines for reading chapters 6 and 7

Given the many accounts there are for the epistemic condition, Wieland (2017, 19–22) mentions two problems that raise suspicions about the possibility of right theory choice to begin with:

Problem 1: intuitions are theory-laden; i.e., theories of different accounts are easily made to fit certain intuitions (or vice versa).

Problem 2: little empirical work has been done regarding the intuitions of non-experts on the epistemic condition (though, see Faraci & Shoemaker 2014). Though, this would not much help if non-experts' intuitions rely on cognitive biases or conceptual confusions (as is likely).

Nevertheless, Wieland (2017, 20) views that any general account of the epistemic condition will need to rely on intuitions about certain cases. Presumably, this is because we humans simply cannot do without intuitions. This appears to be the majority position. However, a number of philosophers encourage us to challenge our intuitions (Rudy-Hiller 2018, ch. 4; see, e.g., Caruso 2018; Levy 2017; Rosen 2004, 296; Zimmerman 2008, 205), though perhaps by appealing to different, conflicting intuitions.

Closely related to these problematizations of right theory choice, it is useful to provide some guidelines for reading the next two chapters that present my own view on the epistemic condition. This seems especially important since the view was mostly formulated before being aware of the broader field of discussion presented in this section (5.4).

Firstly, chapter 6 is primarily formulated in contrast to Sher's views and questioning his intuitions and their consequences. However, what is implicit in the formulation are also possible challenges to some of the other views presented here, and it can be read to some degree evaluating them as well. Secondly, as we will see, it would appear my view has trouble finding its place amongst these dominant views because at least a majority of these are merit-based whereas I am most attracted to a more of a consequentialist view (see sect. 2.2). Thus, my take on how the epistemic condition should be understood seems to be comparatively different.

The following chapter 6 provides a more theoretical answer to the research question of how the epistemic condition *should* be understood, and chapter 7 provides an applied answer. To set the stage, it should yet be emphasized that there are three ways in which "should" can be understood in this context: (1) referring to what is a preferred merit-based view; (2) referring to what is a preferred view in a consequentialist sense (e.g., "we should advocate view X due to its consequential effects, and it may or may not conform with some of our *a priori* intuitions or merit-based views"); and (3) referring to what is a preferred consequentialist view. Due to the shortage of consequentialist views in the last decades (on the area of moral responsibility), what is primarily meant here is the second option. What I'm advocating, and what the "should" in the research question ultimately means, is that we should advocate what I call the pragmatic view due to what seem to be its promising effects for individual, societal, and global well-being, and particularly so in the context of contemporary moral outrage online. Nevertheless, the view will have implications on theory choice among the merit-based views that will be touched on later, especially in sections 6.4.3–6.5 and 6.7n217.

## 6 A FURTHER RESPONSE: THE PRAGMATIC VIEW

In addition to the responses by other philosophers that Sher's account of the epistemic condition has received, his case has also prompted many thoughts of my own. No matter what one's position may ultimately be, his account provides a lot of well-argued and useful nuances for better understanding of the epistemic condition. For example, his nine example cases provide an interesting terrain for examination (sect. 3.3.2), his imaginative reconstruction of the failures to justify the searchlight view provides useful outlines and references for the discussion (3.4.2), and his defense of the normative account of the standard of what an agent should be aware of seems laudable (4.1.2.2). However, despite these commendable features and others, I do primarily have reservations concerning his overall case. In this chapter, I focus on these reservations. At the same time, I build a theoretical answer to the research question of how we should think about the knowledge requirement of moral responsibility. This examination also in part justifies my more applied answer to the research question in chapter 7.

In section 6.1, I introduce my reservations concerning Sher's premise about our common intuitions regarding his example cases and suggest a pragmatic turn in our thinking. In section 6.2, I argue how I think Sher doesn't give enough credit for the importance of paying attention to the engaged perspective. Section 6.3 introduces a thought experiment that illustrates some epistemic problems Sher's account seems to face. Continuing in the pragmatic vein, section 6.4 presents a two-part pragmatic argument for why *not* promoting Sher's view could be seen morally imperative, or at least why putting particular emphasis on the first-person cognitive aspects of the target agent would seem paramount. In section 6.5, I clarify some issues that help to distinguish the pragmatic view from other accounts of the epistemic condition; namely it's deeply normative nature that can be compatible with some descriptive views but not with others. In section 6.6, by introducing and utilizing the 'tragedy of commonsense morality' as a guideline, I locate the approximate home for the pragmatic view among the dominant normative ethical theories. And finally, section 6.7 contains a summarizing definition of the pragmatic view.

### 6.1 Sher's Example Cases and the Problem of Intuition

Echoing also some more or less implicit concerns in some reviews of Sher's book (sect. 5.3), I would like to highlight how Sher seems to be highly motivated in his examination by what he considers to be our common intuitions concerning responsibility (see also sect. 5.4.4.3, 5.4.5). As I introduced in

section 3.3.2, Sher's premise for his suspicion about the deficiency of the searchlight view rests on example cases which he sees to be at odds with it. In the example cases, the agents are unaware of something crucial in their situation, yet still our *intuitions*, according to Sher, would deem them responsible. This goes against the searchlight view as it perceives that agents can only be responsible when at least passively aware of relevant factors in their situation, or – in Zimmerman's qualified version – when their unwitting transgression can be traced to an earlier conscious omission.

Sher's intention was to formulate an account of the knowledge requirement that would better match with how we *actually* intuitively think about responsibility, and with what the searchlight view, according to him, doesn't match with. But there are two questions to be asked here:

- (1) Do our intuitions correspond better with Sher's account than with the searchlight view, in the example cases?
- (2) Why should we trust our intuitions in general, especially if they can be conflicting between individuals?

In sections 6.1.1 and 6.1.3, respectively, I examine both questions, with focus on negative moral responsibility. Section 6.1.2 contains a skeptical interlude where I expand on the initial conclusions of 6.1.1. If Sher's intuitions are as universal as he thinks, the example cases do give credence for the need to expand on the searchlight view in some fashion, for example via Sher's FEC. However, if Sher's intuitions are not as universal as he thinks, then some other conclusion would seem warranted.

### 6.1.1 Our intuitions in Sher's nine example cases

Let us now ask:

- (1) *Do* our intuitions correspond better with Sher's account than with the searchlight view, in the example cases?

In the previous chapter, in sections 5.1–5.3, I presented some reservations raised by Zimmerman (2009), A. M. Smith (2010), Nelkin (2011), Waller (2014), and Talbert (2011). Here, I shall add to those reservations via my own considerations.<sup>105</sup>

Considering my own intuitions in relation to Sher's nine example cases, I can only assure that they are not as straightforward as his intuitions seem to be (see 3.3.2; Sher 2009, 23–29). Of course,

merely stating this gets us nowhere. I shall have to try to untangle how my intuitions seem to differ, on a case-by-case basis.

Before going into the cases, let us briefly refresh our memory of what the searchlight view is. Zimmerman called Sher's basic formulation of the searchlight view as the *unqualified version* of the view, and he concisely summarizes it (2009, 252): the unqualified searchlight view states that "someone is negatively responsible for some act if and only if, (1) he satisfies any non-epistemic condition necessary for such responsibility; and (2) he was [at least passively] aware of the act's wrongness or foolishness." Further, Zimmerman expands the second clause via his *qualified version* of the searchlight view: "(2a) he was aware of the act's wrongness or foolishness or (2b) he was unaware of the act's wrongness or foolishness and this unawareness was a consequence of some other act of whose wrongness or foolishness he was aware." The focus, in this context, is on this overall second clause. Can Sher's example cases be understood via this sort of searchlight view, and what would my intuitions say about the agents' moral responsibility?

In the following sections 6.1.1.1–6.1.1.3, I examine Sher's cases in groups of three that represent the three different types of situations he considers: momentary lapses of judgment, poor judgment, and lack of moral insight, respectively. I listed Sher's example cases in section 3.3.2, which should serve as a reference here, but I shall also briefly summarize the relevant cases as I go through them below. In section 6.1.1.4, I add a regret qualifier on top of my preceding considerations. Section 6.1.1.5 concludes the examination.

#### 6.1.1.1 *Momentary lapses of judgment*

Concerning his example cases 1–3, involving momentary lapses of judgment, Sher says that Alessandra gets distracted and thus forgets her dog, Julian lapses into fantasy and thus fails to notice the rocks, and Wren falls asleep and thus fails to do her military duty. Sher says that in each of these cases the difficulty lies in something overtaking the agent's consciousness, yet they seem intuitively morally responsible. (Sher 2009, 24–25.) I myself do not think the cases are that straightforward.

Sher sees that Alessandra, Julian, and Wren were aware of the wrongness of the relevant act, yet despite their awareness being overtaken and the wrong act occurring while hidden into their unconsciousness, his intuitions would deem them morally responsible. My intuitions, on the other hand, do not seem so sure about such a judgment, and I am not so sure that the *relevant* act is the act we are led to believe here. I think in each of the three cases there can be found reasonable doubt that there may be some awareness concerning *some preceding act's* wrongness or foolishness – implicitly

concerning knowledge – that the agents lacked (even on an unconscious level), and if the agents had not lacked that awareness about *that preceding act* they would have acted differently in the act in question (and they couldn't have acted differently given the lack of awareness in the preceding act). I think it may well be possible that had the agents been aware of certain crucial knowledge that they lacked, their lapses of judgment would likely not have occurred. If there can be found such (lack of) awareness of knowledge, the searchlight view as well as my intuitions would seem to *not* deem the agents morally responsible.

Concerning the case of Alessandra forgetting Sheba in the hot van, my intuitions seem to suggest that she is not morally responsible, given appropriately charitable background assumptions about the case.<sup>106</sup> My intuitions seem to suggest that unless she was aware of how bad an idea of leaving a dog in a hot car in general can be, due to various unpredictable factors that may occur during even the most casual of pick-ups, she is not morally responsible. In fact, if she did lack that awareness, I think after the incident she will have gained relevant knowledge which would then prevent her from making the same mistake again. Essentially, she would have appropriately adjusted the probabilities in her inductive reasoning. In the future, in a similar situation, she would thus always either take the dog with her, leave the dog home or tied outside of the car, or be aware of unpredictable factors interfering upon the safety of the dog in such a way as to be better *metacognitively* conscious of first taking care of the dog when similar unpredictable factors occur. In short, she would have gained important *metacognitive knowledge* (see next note 107 for a definition).<sup>107</sup> However, we may still postulate that there may be some similar lapses of attention afterwards yet, even having learned all of this the first time – but I'll consider this in the next section 6.1.2, after having gone through the cases.

In the case of Julian not noticing the rocks due to his mind wandering, and the case of Wren falling asleep during her military duty, similar reasonable doubt can be raised via my charitable intuition. In both cases there may be some awareness concerning *some preceding act's* wrongness or foolishness that the agents lacked, and if the agents had not lacked that awareness about the preceding act they would have acted differently in the act in question. Like in Alessandra's case, the relevant act here may not be the target of the lapse of attention itself, but a preceding act characterized by a lack of *metacognitive* awareness or knowledge that would have either prevented the lapse of attention or enabled the agent to take necessary (metacognitive or other) *precautions* to eliminate or at least minimize the chance of it happening.

Julian – or, more largely, whoever oversaw the working schedule on the ship – could have been better educated and thus aware of the dangers of mind wandering during routine trips and, if being aware of this, should have, for example, taken precautions to keep additional pair of eyes on the route

in addition to himself, thus paying better attention on this critical segment of the trip. According to my intuitions: if Julian was not aware of the importance of these kinds of precautions, he is not morally responsible for realizing to move the ferry too late – and if he was aware, yet didn't take the necessary precautions, he is morally responsible. Similarly, Wren – or, more largely, the army – may or may not have been aware of the shortcomings of human cognition during exhaustion and night time, and thus, correspondingly, may or may not be morally responsible for not taking the necessary precautions. As in Alessandra's case, here too my intuitions seem to fit together with the searchlight view or something resembling it.

#### 6.1.1.2 *Poor judgment*

Concerning the cases 4–6, involving poor judgment, Sher says that Joliet has poor judgment and thus accidentally shoots her son mistaken as a burglar, Scout has poor judgment and thus ends up causing a child's alcohol poisoning, and Father Poteet has poor judgment and thus causes a traffic accident. Sher says that in each of these cases poor judgment informs the agent's will, yet they seem intuitively morally responsible. Like the first three examples, I don't think these are so straightforward either. (Sher 2009, 26–27.)

Again, I think there can be found reasonable doubt that there may be some awareness concerning *some preceding act's* wrongness or foolishness that the agents lacked, and if the agents had not lacked that awareness about the preceding act they would have acted differently in the act in question. And my intuitions, like the searchlight view, seem to say that if there could be found something lacking in their knowledge, the agents are not morally responsible.

I think Joliet is not morally responsible if she was not properly aware of her clearly severe psychological phobia being able to cause such harsh reactions as to making it an imperative, as a precaution, to either emphasize to all family members not to sneak up to the house uninformed or to eradicate all guns from the house.<sup>108</sup> Similarly, Scout, my intuitions say, is only morally responsible if she was either aware that alcohol and children do not mix well medically or if she was aware why it would be important to look up information on whether that is the case (if she was not aware of either of these, she is not morally responsible). And, finally, Father Poteet's split-second decision could be either considered a choice between two or more *possibly* bad choices (see Aristotle's mixed actions in section 2.1.1), the outcome of which he could not have predicted, or lack of relevant knowledge concerning the kind of split-second choice he was forced to make, and thus he would not be intuitively morally responsible. In all these cases, and the others, the agents may be legally

responsible, as well as of course causally responsible (as part of the causal events leading to the outcome), but I would not necessarily deem them morally responsible. Thus, here too, my intuitions would seem to be in alignment with something resembling the searchlight view.

#### 6.1.1.3 *Lack of moral insight*

Concerning the final three cases 7–9, involving lack of moral insight or imagination, Sher says that Ryland is too self-centered and thus fails to recognize the impact of her anecdote, Sylvain is too focused on the individual before him and thus fails being fair to others, and amerika has taken a wrong turn in working out the implications of his moral beliefs and thus commits a crime. Sher says that in each of these cases the agent simply lacks certain moral insight or imagination yet seems intuitively responsible. As before, I'm not so sure these cases are that straightforward. (Sher 2009, 25–27.)

These cases, however, are a bit different than the preceding ones. In the case of Sylvain, I think there can be found reasonable doubt that he lacks some crucial knowledge, which, if he did not lack, he would act differently in the acts in question. And, again, my intuitions, like the searchlight view, seem to say that if there can be found something crucially lacking in an agent's level of knowledge, the agent is not morally responsible. The cases of Ryland and amerika are a bit different still, which I'll explain below as I go through the cases.

In the case of Ryland, I think her anecdote may well have been poorly received by the audience, but I am not at all sure if we should consider people morally responsible merely for some bad jokes or other words being subjectively distasteful to some in the first place (whether or not the speaker or writer is aware of the audience including members who might find them distasteful). Nor am I sure we should consider people responsible for their failure to predict what a group of people, or some individuals within a group, might think in response to their jokes – i.e., for their failure to predict what kind of subjective sense of humor or what sensibilities an audience may have.<sup>109</sup> I think a conclusion on this part would require specific arguments concerning the limits of freedom of speech (e.g., how should we interpret the *harm principle*<sup>110</sup>), and a more specific characterization of what exactly it was that Ryland said. In this context, we may *assume* that Ryland somehow violated the limits of freedom of speech in such a way that I would be willing to accept there being moral responsibility, given that Ryland had the relevant preceding awareness or knowledge. But what would be relevant in this case would then be the arguments concerning the limits of freedom of speech and relevant consequences that Ryland should be aware of before I would be willing to consider her morally responsible. It could also be that Ryland has a well justified position on freedom of speech

and how that relates to moral responsibility, and arguments for her position that Sher nor I may not have heard of (in which case we cannot be held responsible for not being aware of those arguments, and Ryland would do well to share them).<sup>111</sup>

As for Sylvain, I would not consider him morally responsible for being unfair to his students unless he was aware of what negative effects too much *emotional empathy* may cause, and do cause, in his case. Specifically, I would recommend Sylvain to read psychologist Paul Bloom's arguments and related studies in his book *Against Empathy: The Case for Rational Compassion* (2016). If Sylvain had read Bloom's book, or something similar, and reflected his life via the arguments and research concerning the downsides of emotional empathy, and afterwards continuing to act how he does, then I would be willing to deem him morally responsible – for he would then seem to deliberately neglect the ethically important knowledge he has acquired. After having properly read the book, Sylvain would seem to have all the (metacognitive) knowledge he would need to become aware of his situation and what a relevant self-intervention might look like via precautions that take note of the downsides of emotional empathy.<sup>112</sup>

Finally, I think the case of anti-capitalistic amerika and robbing the bank is, in a way, closely related to the case of Ryland and her bad joke. I think the question here is also more so concerning what should qualify as moral principles or their limits that guide our actions and ideologies, not how moral responsibility should be dealt with *per se*. If amerika was clearly involved in killing the guard, he should be held responsible, both legally and morally, but in the latter sense only given that he was aware of what entails a possible killing or other physical harm and why avoiding such a possibility would be ethically more important than his radical ideology. It *could* also be that amerika has a superior justification for his ideology that overrides any attempts of justifying not committing a crime (even though we may intuitively have trouble imagining this; but we have not heard his argument).

#### 6.1.1.4 *The regret qualifier*

To add one further criteria to the cases that I did not mention above but merely seem to have intuitively assumed to have been met: Taking note from Aristotle (sect. 2.1), some of the agents that otherwise seem to have acted without their respective awareness ought to also exhibit symptoms of feeling other-regarding sorrow and regret after becoming aware of their preceding unawareness, for them to qualify as pardonable.<sup>113</sup> This would be to try to rule out the agents only pretending to have done the act without their awareness of what they were doing. For example, if Joliet pleaded to unawareness for having shot her son (in the way described above), and for us to thus pardon her, yet she does not

feel other-regarding sorrow and regret but instead seems to feel neutral or positive about the event, this would give us good grounds to believe that she might be bluffing to have made a mistake without awareness or, alternatively, that her threshold to shooting her son was so low to begin with that her initial unawareness of the victim's identity played a practically inconsequential role in preventing her from shooting her son. The same criteria would similarly apply to all the cases except possibly the cases of Father Poteet's split-second decision, Ryland's bad joke, and amerika's crime.<sup>114</sup>

The problem in applying the regret qualifier to Father Poteet seems to be that *if* he was forced to choose between two or more bad decisions, I would not necessarily expect him to feel any kind of sorrow and regret for choosing something in a situation in which positive choice was impossible. It would seem that in such a scenario he would have been acting in compulsion, thus not fulfilling the voluntariness condition (i.e., he was forced to choose between two equally bad choices, which was out of his control). However, *if* the decision truly lacked some crucial knowledge concerning the kind of split-second decision he was forced to make, then the regret qualifier would apply to him as well.

In the cases of Ryland's bad joke and amerika's crime, the problem in applying the regret qualifier is that *if* they do have a superior justification for their acts that we are not aware of, then those who have blamed them without hearing them are the ones who ought to feel other-regarding sorrow and regret. However, if opposite arguments to theirs' are the superior ones, then the regret qualifier would apply to the agents (after they have learned about the arguments). It might also be that the moral argument of neither side is superior but merely equal or roughly equal in their strengths and weaknesses, in which case there would not even be one right answer to whether their acts were properly justified.

#### *6.1.1.5 Conclusion – Agreeing to disagree on metacognition*

Having thus laid out my interpretations of Sher's example cases and my corresponding intuitions of the agents' responsibility, it seems that the cases *can* be understood via a view at least resembling the searchlight view, and, furthermore, my intuitions seem to conform with it. Thus, Sher's intuitions about the example cases do not seem universal. Hence, his expansion of the searchlight view – at least when grounded on these example cases – can only serve particular intuitions that not everyone shares.

It seems likely that Sher would maintain that the agents remain intuitively morally responsible in the cases. If not for the acts he suggests them seeming responsible for, then for the preceding acts or omissions behind their failure to take necessary precautions, or not having actively acquired the

preceding and needed relevant knowledge or awareness, be it metacognitive or some other kind. Thus, I suspect he may think that even if the agents were not intuitively responsible for the acts in question, then at the very least they were responsible for their lack of preceding metacognitive knowledge, or for their lack of research in their situation or beforehand. He might add that thus, by extension, the agents are also indirectly responsible for the acts in question. Further, their failures might be caused by their *constitutive features*, which Sher would no doubt appeal to in support for his intuitions. This is a fair position to take in the sense that, indeed, it seems our intuitions vary.

It may merely be the case that our subjective intuitions lead us to different conclusions on individual (or group) basis, even after all the possible nuances, i.e. background assumptions, of the overall situations are agreed upon. In other words, we may intuitively put different emphasis on the importance of being aware of something, or knowing something, and how that relates or should relate to deeming an agent morally responsible.

In the example cases, taken as they are, it seems my intuitions pay particular focus on the metacognitive ability of the agents and whether or not they had gained relevant knowledge – metacognitive or otherwise – that would have enabled them to take reasonable precautions to prevent the unfortunate faults in their (meta)cognitive abilities that led to the acts Sher would seem to want to hold the agents *de facto* responsible for. Further, it seems that acquiring new metacognitive knowledge can be thought in terms of enhanced *searchlight control* (i.e., enhanced ability to track relevant facts in one’s consciousness).<sup>115</sup>

I have no trouble admitting that I may be in a minority with my intuitions. I merely hope to have shown how our intuitions may in fact conflict, on an individual level, when examining Sher’s example cases. Thus, we may be intuitively led to different conclusions on the question of whether Sher’s example cases demonstrate the responsibility of the agents, and whether they demonstrate problems of the searchlight view in the way he intends. To answer the first question proposed at the beginning of this chapter: our intuitions *do not* necessarily correspond better with Sher’s account than with the searchlight view, in the example cases.<sup>116</sup>

### 6.1.2 An interlude of doubt

Despite my intuitions seeming to fit together with something resembling the searchlight view, there is a nagging question at the back of my head. Namely, are my interpretations of the cases merely shifting the problem Sher is trying to outline to a possible future act (or omission)? For example, in a sense, I am shifting the relevant act from *Alessandra forgetting her dog* to Alessandra not being

metacognitively knowledgeable enough to take the *necessary (metacognitive) precautions* to prevent her from accidentally allowing Sheba to languish. It seems to me that Sher could say that *after* Alessandra has gained the metacognitive knowledge, and *if she then* experiences a *metacognitive* lapse in a reoccurring similar situation, despite having previously acquired the metacognitive knowledge that in most cases would guard Sheba, and Sheba is yet again forgotten, we are back at the original problem. In other words: yes, my intuitions can fit together with something resembling the searchlight view in the example cases *as they are*, but if we were to add to the cases that the agents' failure happened after them having previously gained the kind of metacognitive or other knowledge that I suggested they lacked, how would that change my intuitions fitting together with the searchlight view?

For example, let us assume that Alessandra had already once forgotten Sheba in the hot van, and had effectively learned a metacognitive lesson and started to follow a careful heuristic: during warm days, always tie Sheba to the car outside and/or always pay particular attention to the thought of Sheba if unpredictable factors occur while picking up children from the school. This works great, and Sheba is happily cared for every time Alessandra goes to pick up her children. Until one day, when Alessandra is in a hurry, she leaves Sheba in the van, goes to the school and encounters another situation that suddenly requires her careful attention. Despite having developed the careful heuristic, and having acquired the metacognitive knowledge, she experiences a rare cognitive lapse and once again forgets Sheba (and this is caused by the interaction of Alessandra's *constitutive features*, one way or another).

How would this kind of a scenario fit together with the searchlight view? The plausibility of the scenario, in the first place, hinges on the actual possibility of these kinds of scenarios happening without serious cognitive defects that would be more broadly considered pardonable. Having considered the example cases again via assuming all relevant metacognitive knowledge to have been gained beforehand, I find it to be *unlikely* for the agents to fail afterwards in the required ways to enable scenarios where there could be even considered to be responsibility without awareness (i.e., scenarios where the agents would not be even passively aware of the relevant (meta)knowledge, after having come to possess, consolidate, and utilize it earlier).<sup>117</sup> This puts the possibility of these kinds of scenarios into doubt. However, I cannot completely rule them out.<sup>118</sup> The kinds of scenarios in question are theoretically possible but seem practically improbable. This leaves the possibility, however narrow, for there to still be cases that may not intuitively conform with the searchlight view, given intuitions that hold the metacognitively reinforced Alessandra responsible. The searchlight

view itself would not hold Alessandra responsible in this case any more than in the original, as she was deprived of crucial awareness in both.

And how would my intuitions react to these kinds of metacognitively reinforced scenarios? Insofar as these scenarios can happen, I do find myself still doubting if I would hold the agents morally responsible even then – given that they feel genuine other-regarding sorrow and regret afterwards, feeling that they could not help their cognitive lapse despite their precautions, including the previously acquired metacognitive precautions. Thus, in the metacognitively reinforced cases my intuitions would also seem to conform with the searchlight view or something resembling it. Presumably, the agent’s cognitive lapse would yet again reinforce their relevant (meta)cognitive processes, albeit they may still fail under some future circumstances, however even-more-unlikely it may then be.

Herein, I think, lies the most considerable departure between Sher and me: at least in this context, I find the *control condition* extremely intuitive and appealing. If the agents were not in control of the tricks of their mind that led them to falter even in the light of their previously acquired metacognitive knowledge, I find it hard to blame them any more than they are likely to already blame themselves for factors outside of their voluntary control. In fact, it seems highly *uncompassionate* to blame them. Alas, there is a lurking road of *moral responsibility skepticism* that looms if this direction is thoroughly embarked upon.<sup>119</sup> After all, the deliberately vile act of a violent criminal seems to be a product of events – determined or indetermined – that are as much outside of his control as the events that produce the most altruistic benefactor in history are of hers (see sect. 2.2; see also Caruso 2018, ch. 2). Still, I find it intuitive to hold agents morally responsible if they are deliberately doing vile acts, or if they are deliberately neglecting metacognitive heuristics they have formed after coming to learn to form them for good reasons. However, to blame them or to forgive them, I am not sure how to judge *my intuitions*.

### 6.1.3 The problem with intuition

Given the apparent failure of our intuitions, in the sense that they can conflict, and further noting that I have hard time conciliating even some of the varied nuances of my *own* intuitions, we may move on to the second question I presented at the beginning of section 6.1:

- (2) Why should we trust our intuitions in general, especially if they can be conflicting between individuals?

Of course, this is largely a rhetorical question. We should not automatically just trust our intuitions, especially if they can conflict between individuals, let alone within an individual.<sup>120</sup> Our intuitions concerning the role of knowledge or awareness in moral responsibility is hardly an exception in the sense that our intuitions seem to be at odds at times. As it happens, even Sher wrote that he sees the searchlight view as “the default position to which we gravitate when we are not thinking hard about the knowledge requirement” (2009, 7), yet, ironically, thinks his intuitions reliably guide us *away* from the searchlight view (cf. sect. 3.2).

Below, I provide an illustration of what I see to be the basic problem with intuition, and how I think it encourages us to consider a pragmatic angle in our approach to the epistemic condition (6.1.3.1). Further, I consider our current practices that are presumably to some significant degree guided by our prevailing intuitions about responsibility (6.1.3.2).

#### 6.1.3.1 *The basic problem and a pragmatic turn*

The basic problem with intuition can more generally be characterized as the often-made observation in our everyday lives – as well as in the history of false but intuitively appealing knowledge claims<sup>121</sup> – that our intuitions can lead us astray and that there can be conflicting intuitions both between and within individuals as well as groups of people. For a somewhat relevant example for this context, I have encountered people who claim that their intuitions support a harsh retributive or deterrent system of punishment as a means to a more peaceful society, whereas some other people claim that their intuitions say a more restorative or rehabilitative approach to justice would yield better results.<sup>122</sup> As a more commonplace example, we may think how our intuitions concerning what kind of nutrition (or diet) would be the best for our health goals can conflict between individuals, and further be in conflict in relation to (research about) what nutrition would *actually* benefit us the most.<sup>123</sup> In just about any branch of science there is almost a hyperawareness of just how mistaken our common intuitions can be, and thus research to test our intuitions via testing out hypotheses is valued. Relatively often, research can reveal results that undermine our prior intuitions.<sup>124</sup> Therefore, even though if the current intuitions of many people, even the majority, suggest one thing as being the case, the case may in fact be something different entirely.

Concerning especially moral intuitions, we may further consider their often, if not always, quickly reacting and emotion-laden nature. We often feel strong and uncomfortable feelings of *cognitive dissonance* whenever our morally implicated intuitions about, for example, justice or nutrition are deeply challenged, and even if this happens on strongly evidential grounds.<sup>125</sup> Likewise,

those who do not share our intuitions often feel just as uncomfortable when we challenge their sentiments, even if on evidential grounds. To paraphrase social psychologist Thomas Gilovich (1991, 83–84): when we want to believe something, we effectively ask “*can* I believe it?” (i.e., can I find *any reason to believe* the *a priori* conclusion?); and when we don’t want to believe something, we ask “*must* I believe it?” (i.e., can I find *any reason to doubt* the *proposed* conclusion?) (see also Ditto & Lopez 1992). The standards of evidence for the two questions are quite different. It seems to be a comparative rarity, if not impossibility, to emulate a neutral observer and merely ask “what does the *overall evidence* suggest?”.

However, when we do manage to tame our passions – however difficult it may be – it seems easier to find approaches that also tame these kinds of communication dilemmas and help us see the cumulative evidence behind or outside them, no matter what direction the evidence may ultimately point to. Taming our passions may also help us generate more trustworthy intuitions (Stanovich 2018a). Still, we appear to remain highly prone to see what our initial passions want to see whenever there is any ambiguity involved (e.g., Balcetis & Dunning 2006). More specifically, we remain prone to *confirmation bias* (e.g., Blanco & Matute 2018; Hart et al. 2009; Nickerson 1998), *motivated reasoning* (e.g., Kunda 1990; see also Balcetis & Dunning 2006; Ditto & Lopez 1992; Ditto, Pizarro, & Tannenbaum 2009; Druckman & McGrath 2019; Nir 2011), finding *ad hoc explanations*, and to many other cognitive blunders, distortions, and biases that can compromise our understanding of reality and, as a consequence, our functioning in accordance with it (for illustrative lists, see Novella 2018, 44–140; Pohl 2017; see also Gilovich 1991; Haidt 2012, 84–108; Kahneman 2011; Nisbett 2016; Sutherland 1992/2013; see also sect. 6.4.2).<sup>126</sup> Moreover, we seem to be particularly motivated by the views that appear to be held within our perceived in-group (e.g., Balliet, Wu, & De Dreu 2014; Ditto et al. 2019b; Thürmer & McCrea 2018; see also Delamater et al. 2015, 452–460 & 481–485; Taber & Lodge 2006; sect. 1.1 & 1.2; see also intergroup attributional bias: sect. 6.2.1n137), and especially so in polarized settings (Druckman et al. 2013).<sup>127</sup> At the same time, we are prone to exhibit *bias blind spot*: to naïvely perceive we are less biased than most others (e.g., Pronin, Gilovich, & Ross 2004; Pronin, Lin, & Ross 2002; Scopelliti et al. 2015), a result recently replicated in samples from the US and Hong Kong (Chandrashekar et al. 2019). Thus, all things considered, caution remains warranted whenever we are dealing with intuitions, or reasoning that heavily relies on them.<sup>128</sup>

This sort of description can be viewed as something approaching a Humean take on the nature of intuitions, and similar views have been widely advocated by moral psychologists and philosophers, and others, on both rational and empirical grounds (see, e.g., Cushman, Young, & Greene 2010; Ditto et al. 2009; Flanagan & Richardson 2010; Greene 2013; Haidt 2001, 2011b, 2012; Kahneman 2011;

Kauppinen 2013; see also Aristotle & Reeve 2014, X.9.1179b22–28; Hume 1739–40, T 1.3.1.2, SBN 70; T 2.3.3.4, SBN 414-5; sect. 4.1.4n81). Not all nuances are agreed upon, however, but the general theme of there being two general types of thinking processes seems descriptive of these views. Nowadays, these commonly build on or refer to various versions of what are known as *dual-process theories* of higher cognition, in which context the general distinction between our automatic, intuitive, fast-responding evaluations and reflective, analytic, slow deliberations are referred to as the distinction between Type 1 and Type 2 processing, respectively; also called System 1 and System 2 (Evans & Stanovich 2013; see also De Neys 2018; Kahneman 2011).<sup>129</sup> Relating to the previously mentioned importance of metacognition (sect. 6.1.1.1n107), variation in the functioning of these two types of thinking seem to be connected to variation in the efficiency of our metacognitive processes via those processes determining when we engage in Type 2 reasoning (Thompson 2009; Thompson, Turner, & Pennycook 2011; Thompson, Evans, & Campbell 2013).<sup>130</sup> In effect, it seems that the more dogmatic our intuitive judgments are, on a particular case, the less we are prone to do so (Thompson et al. 2011, 2013; see also note 125).<sup>131</sup>

It may further be described that all perceptions are at least potential vehicles for arousing our emotions and bringing about feeling states via Type 1 processing. Yet, feelings that a shared perception can bring about at a given time can vary individually. For example, even though I tend to feel inspiration and hopeful optimism when watching the fictional series *Star Trek* (particularly *The Original Series* and *The Next Generation*), someone else may tend to feel dismay instead, and yet a third person may feel apathy or boredom, and so on. This variation is most likely explained by some combination of environmental and biological/neurological variation in our personal histories (including our upbringing, family histories, cultural histories, and epigenetic, genetic and evolutionary histories – along with related epistemic histories, histories of imagination, histories of illness or health, and histories of bias inoculation, etc.). Our upcoming histories may of course further affect our views. It seems that most, if not all, intuitions work in this way. The same stimulus X – for example, the example of Alessandra forgetting her dog in the van – can bring about different intuitions (of moral judgment) in different people at a given time (and emotions, which most likely follow those intuitions, or vice versa, or they may be co-constituted, or loop in various ways; see Avramova & Inbar 2013). There are most likely some biological species-specific boundaries to what intuitions a specific stimulus can induce in a body experiencing consciousness, but within those limits, and further based on our individual histories, intuitions appear to vary. And this would appear to apply to philosophers as well (Sinnott-Armstrong, Young, & Cushman 2010, 268–270; see also sect. 2.2n50).

Especially since our intuitions in Sher's example cases would appear to vary, and thus as our approach to moral responsibility would appear to be not constant across individuals, I would be prone to think about the matter more pragmatically: Given that our intuitions vary to one degree or another, and given that our relevant intuitions are and can be shaped by our upbringing and other environmental factors (e.g., Aristotle & Reeve 2014, X.9; Sher 2001; see also sect. 2.1n33), it would be productive to approach the question of our intuitions about the epistemic condition via a question of what kind of thinking about the matter would be most conducive to a well-functioning coexistence within a society. In other words: how *should* we think about the matter – for example, in terms of best guiding people's future actions, and/or upholding individual, societal, and global mental health (within the boundaries that biology permits)? Or, more broadly, how should we think about the matter in terms of cultivating more widespread *eudaimonia*; that is, human flourishing (see, e.g., Flanagan 2011, 95 & 158–159 & 201–202)? Later, when the answers would be approached, we could then focus on trying to encourage that sort of thinking on a wider societal level, via our institutions and general educational work, so that in time we could be molded to better follow a form of moral evaluation and judgment that would be more uniformly conducive to individual, societal, and global well-being. The related cognitive processes can be noted to relate to moral *metaevaluation*, *metajudgment*, or *metaresponsibility*<sup>132</sup>: the (study of) monitoring and control of our evaluations, judgments, and conventions relating to moral responsibility.

This pragmatic turn can be seen as a shift from the merit-based view to the consequentialist view regarding moral responsibility, where the latter has been largely neglected in the last few decades (sect. 2.2). Even though it seems to be a debatable question of which direction Aristotle himself advocated, this is nevertheless the direction I think ought to be taken. At the same time, we may read Aristotle as talking about this direction (see ch. 2), and, furthermore, we may add to and enhance his thoughts via the accumulating empirical evidence that is nowadays available to us.

#### 6.1.3.2 *Our current practices*

In addition to Sher claiming that our intuitions would (universally) not match with the searchlight view, he also says that our actual practices do not match with it (see sect. 3.3.2). Given that our common practices are largely guided by our intuitions, we may find some corresponding individual variation there as well. I would imagine my practices largely corresponding with my intuitions (in Sher's example cases), at least insofar as they are not clouded by some distorting cognitive biases or emotions in particular circumstances, and insofar as I am not coerced to act against them. However,

as Zimmerman can be seen to note (2009, 254), Sher's account may very well capture the everyday moral judgments *of the majority*, and thus the most common practices concerning moral judgment. It may actually be the case that our current most common individual intuitions and consequent practices correspond to Sher's account. But I don't think there are good reasons to yield the matter to the hands of the current possible majority intuitions and practices.

In fact, I think the current possible majority intuitions and practices may be something we should be very well worried about. In recent years, this concern is well demonstrated, for example, on social media. In sections 1.1 and 1.2, motivating this examination, I summarized some of the research concerning moral outrage online, and of the related peculiarities of human interaction. Those phenomena may partly be seen as manifestations of these kinds of individually vaguely defined and fuzzy intuitions about responsibility running wild, largely ignoring or being unaware of the epistemic component in responsibility, properly understood. If nothing else, the political polarization and seemingly commonplace hasty knee jerk reactions, moral judgments, and modern-day witch hunts between various polarized groups are certainly *not* signs of calmly thinking things through.

Granted, Sher thinks awareness does still matter to some significant degree (via his clause 1 of FEC), and he may be well motivated to communicate this to a public that seems misguided when judging each other. However, I do not think he gives enough credit for it mattering – especially since [A] our intuitions in his example cases do not match; and since [B] he seems much quicker to blame and assign negative moral responsibility to agents, and even punishment (see sect. 3.3.2; cf. 6.1.1); and because [C] he pays no explicit attention to metacognition (see sect. 6.1.1.1n107; 6.1.1.5). More in line with other thinkers (see ch. 5), I think the role of (meta)knowledge or (meta)awareness in moral responsibility should particularly be emphasized, if not fully endorsed, via some variety of the searchlight view or, perhaps, Smith's attributionism (or, perhaps, via some view that is not much examined in this thesis).

I think there are good *pragmatic* reasons for emphasizing the role of (meta)knowledge in agent evaluation, and I present an explicit pragmatic argument for those reasons in section 6.4. But before that, in section 6.2, I raise some relevant reservations I have about the issue regarding the engaged and the detached perspective; and, in section 6.3, I examine some epistemic challenges Sher's account would seem to face.

## 6.2 On Why the Engaged Perspective Should Always Matter

While the searchlight view fails to match with Sher's intuitions, he sees that the fundamental problem with it is that it conflates two different perspectives on action: the engaged perspective and the detached perspective – as introduced in section 3.3.1, and particularly laid out in 3.4. As also emphasized in section 3.4, Sher is adamant that there are no considerations that would compel us to ignore all cognitive states to which the agent lacked access to. This he takes to be contrary to the searchlight view and hence to support the deficiency of it; and to support how the natural home of the concept of responsibility is the detached perspective, not the engaged perspective nor solely taking note of the engaged perspective via the detached perspective – a view his intuitions, as demonstrated via his take on the nine example cases, also support.

Concerning the distinction between the engaged and the detached perspectives, and the question of where the natural home of responsibility resides, a critical question regarding Sher's account can be proposed: Is it *truly* the case that the natural home of the concept of responsibility is the detached perspective? I think this is not necessarily the case. In this section, I explain why I think so.

In section 6.2.1, I illustrate how the two perspectives seem to be interlinked in ways that encourage us to pay careful attention to the engaged perspective. In section 6.2.2, I distinguish between being responsible and taking responsibility, which also encourages attention towards the engaged perspective. And section 6.2.3 briefly concludes this examination while also introducing the concept of “epistemic state”.

### 6.2.1 Interlinkedness of the engaged and the detached perspectives

To begin, it is useful to note that Sher is adamant that these two perspectives cannot be occupied at the same time. In one sense, this does seem to be the case – but in another, it seems not to be the case. On the one hand, Sher seems to be right: when we perform an action, our perspective is one engaged in the act, and only after the act can we occupy a genuinely detached perspective in relation to the act. But, on the other hand, when we are occupying the detached perspective, from which we consider the past action, that in itself can be seen as an act (of consideration) we are engaged in, albeit a different act than the past act we are considering.<sup>133</sup> This seems to be an important distinction when wanting to understand the value of separating the two perspectives from each other: they are a valid dichotomy but nevertheless the engaged perspective is always present, in one way or another. For every detached perspective concerning a given past act, there seem to be at least three distinctly

relevant engaged perspectives: (1) the engaged perspective that was occupied during the past act that is now being considered from the detached perspective, and which we cannot occupy simultaneously; and (2) the engaged perspective that concerns the act of considering the past act from the detached perspective, and which we always occupy simultaneously; and (3) the timeline of engaged *perspectives* that have concerned the acts that lead an agent from either occupying the engaged perspective during the past act (if the agent is the actor) or from having spectated or heard about the past act happening (if the agent is an outsider) to occupying the engaged perspective concerning the act of considering the past act from the detached perspective. Clearly, Sher is only referring to the first kind of engaged perspective, but, while doing so, seems to fail to notice its overarching importance. Below, I will be referring to the engaged perspective in a general sense, to illustrate this importance.

For us to properly engage in the detached perspective after an act – in ways where we can confidently say something blame- or praiseworthy was done (regardless of our thesis of the epistemic condition) – some *a posteriori* knowledge would seem to be required (in any case, but especially in Sher's nine example cases that can be interpreted with a variety of background assumptions, as demonstrated in ch. 5 & sect. 6.1.1). In other words, for us to get to talk about the engaged or the detached perspective after the act, some new knowledge – or at the very least information of some relevant kind – needs to be fed into the cognitive system via the engaged perspective (be it the cognitive system of the agent whose actions are in question or the outside agents who are evaluating the responsibility of the agent). This is something that Sher doesn't seem to address at all.

Failing to address this, it seems that Sher is belittling or too harshly dismissing the value of the engaged perspective. As the detached perspective requires knowledge that can cognitively only be obtained via the engaged perspective, respecting the engaged perspective (on the part of anyone who is feeding new information to other cognitive systems) is in a key position whenever we would like to bring about appropriate reflection from the detached perspective, concerning the actions of the agent or others. Thus, it seems crucial to *always* pay attention to the engaged perspective – and this is something that is further emphasized by my more illustrative pragmatic argument that I present in section 6.4 (and section 6.4.2 especially). Consequently, the natural home of responsibility doesn't seem to be exclusively the property of the detached perspective, nor the engaged perspective, because the perspectives are interlinked.<sup>134</sup> There is no detached perspective without the engaged perspective.<sup>135</sup>

This kind of justification to emphasize the engaged perspective doesn't seem to *strictly* fall under any of the considerations that Sher engages with to find a reason to ignore all cognitive states

to which the agent lacked access to during the act. As presented in section 3.4, Sher's considerations were about (1) the practical nature of the concept of responsibility, (2) the unfairness of blaming or punishing people for what they cannot help, (3) the connection between imaginatively identifying with an agent and understanding what s/he has done, and (4) the unreasonableness of any demand with which the agent to whom it is directed is unable to comply (Sher 2009, 11). It could be characterized that the consideration I have presented concerns the interlinked quality of the two perspectives, and thus arising pragmatic importance of always paying attention to the engaged perspective.<sup>136</sup> Still, this does not necessarily compel us to ignore *all* cognitive states that the agent lacked access to during the act, but it does compel us to *not* ignore the perpetual cognitive importance of the engaged perspective. Essentially, while taking note of the cognitive state of the agent, this shifts the focus to other agents and how they should note the cognitive state of the agent whose acts are in question (and how they should (metacognitively) note their own cognitive state while evaluating acts of others or themselves). The engaged perspective should always be noted if we want to optimally bring about appropriate reflection, or other cognitive processes, in other subjects of the world or in ourselves. Also, it helps us to understand ourselves and others better.<sup>137</sup>

### 6.2.2 Being responsible versus taking responsibility

The natural home of responsibility not necessarily being the detached perspective is also noticed when distinguishing between someone (a) *being responsible* from someone (b) *taking responsibility*.<sup>138</sup> The former refers to a backward-looking view (or a present-looking view if the act is ongoing), whereas the latter refers to a forward-looking view. That is, someone *being* responsible for a past (or occurring) act is an evaluation made from the detached perspective (and often from a detached perspective of someone who is not the target agent), while *taking* responsibility for future acts is something that happens from the engaged perspective of the target agent. Thus, while the former can involve both the engaged and the detached perspectives in the interlinked manner described in the previous section (6.2.1), the latter seems to at least directly only involve the engaged perspective (indirectly, taking responsibility can be a conclusion reached via engaging with the detached perspective). Hence, to best enable an agent to *take* responsibility, paying attention to the engaged perspective is also called for.

Following from this distinction is a corresponding distinction of what “*accepting responsibility*” can mean. It can mean either (a) accepting *being* responsible for some past act, from the (engaged) detached perspective; or (b) accepting responsibility in the sense of *taking*

responsibility, from the engaged perspective. Thus, it can be understood that when an agent ‘accepts responsibility’, what they are accepting is not necessarily concerning being responsible for some past act but it may merely be an admission of new (meta)knowledge having been learned that they are now aware of and willing to henceforth incorporate into their decision-making processes lest they *otherwise become* responsible (given no further excuses), and liable to be held responsible for a similar future past act, for now they have come to possess the requisite (meta)knowledge (given no further involuntary cognitive lapses). Further, *holding* an agent responsible can also take two corresponding meanings: either the agent is held responsible in the sense that (a) they are held responsible for (*being* responsible of) a past act, or (b) they are held responsible to henceforth *take* responsibility. Similar dual meaning applies to *assigning* responsibility, and other possible related terms.

Of course, in the cases of accepting, holding, and assigning responsibility both meanings (a) and (b) can be referred to at the same time – and it may consequently be that they are often confounded. Further, it seems (a) usually implies (b), but not vice versa.

For the purposes of the remaining thesis, it is helpful to emphasize that these two meanings are closely connected, but seem rarely explicitly distinguished in the literature. This is the case even in this thesis, though it is contextually deducible which meaning is being referred to or whether both are. They are closely connected via the same relevant (meta)knowledge, in a particular case, being involved in enabling both. For example, Alessandra having (or not having) (meta)knowledge concerning leaving a dog in a car is involved in both her being (or not being) responsible for the past act and her taking (or not taking) responsibility for similar future acts. And, indeed, it seems to be implied that *if* Alessandra is responsible in the former sense, she is also responsible in the latter, but not necessarily vice versa.

### 6.2.3 Conclusion and epistemic states of agents

Overall, the conclusion here is not against Sher’s in the sense that it would support the notion that holding an agent responsible requires regarding what s/he did *exclusively* from his or her own perspective (via our third-person perspective). That doesn’t seem to be the case necessarily. But, at the same time, it does not agree with Sher in the sense that even though the agent’s perspective should not be regarded *exclusively*, it still should always be regarded – and regarded in a pragmatic sense.

The engaged perspective of the target agent should always be regarded in such a way as to lead to well justified or otherwise properly action-guiding considerations from their detached perspective,

which, because the detached perspective depends on the engaged perspective, requires us to carefully evaluate the cognitive state of the agent both during and after the act as well as the cognitive state of those evaluating the act from the outside (6.2.1).<sup>139</sup> This evaluation of the cognitive states of the participants is further encouraged by the target agent taking responsibility relying on the engaged perspective (6.2.2). Evaluating the cognitive states of the evaluators is encouraged for the purpose of making sure they are properly engaged in the evaluation, by them appropriately noting the engaged perspective of the target agent.

To focus on the especially relevant aspects of the cognitive states that should always be evaluated, we may specifically want to focus on the *epistemic states* of the agents. By “epistemic state”, I refer to an individual’s set of conscious and unconscious beliefs, (meta)knowledge, and other (meta)cognitive epistemic properties at a given time, at least partly based on which the individual makes his or her epistemic and moral judgments that shape their actions (see sect. 6.1.3.1).<sup>140</sup> In adult population, these cognitive properties are relatively stable but still malleable in propitious contexts, which contexts involve new information being successfully fed into the cognitive system of the individual. That is, the contexts involve successful belief or habit revision to induce belief or habit change in the target agent, significantly changing their epistemic state. In the following two sections (6.3 and 6.4), this will be my implicit focus.

### 6.3 Epistemic Considerations: The Case of Bob and Global Warming

In this section, I illustrate via a thought experiment why the importance of a knowledge requirement in moral responsibility should not be underestimated.<sup>141</sup> To make my thought experiment topical, it utilizes climate change as an example.

Amongst those who read this thesis, I would assume the urgency and very high scientific probability of anthropogenic global warming (henceforth abbr. AGW) to be a relatively non-controversial issue (see also sect. 1.2). It should therefore function as a suitable example. However, if someone disagrees with the state of evidence, for my argument they should assume that the state of evidence is clear.<sup>142</sup> I also take it to be relatively non-controversial that global warming should be mitigated, or, ideally, stopped. My aim is *not* to argue the science but to argue via an example I take to be relatively uncontroversial in terms of an agent being misguided. The aim *is* to accentuate the knowledge or (un)awareness dimension, in order to get a view from which to examine the knowledge requirement specifically. Furthermore, the agent in the thought experiment may be noted to induce

outrage in some people, in some contexts, particularly in some groups on social media (especially as those outraged would usually know little of his background). In section 6.3.1, I introduce the thought experiment. In sections 6.3.2 and 6.3.3, I illustrate two epistemic concerns it raises.

### 6.3.1 Introducing Bob and two views to his responsibility

To begin, let us imagine a moral agent whose name is *Bob*.<sup>143</sup> Bob is a Texan male, living in an individualistic, free market culture, surrounded by pertinent religious fundamentalists, and Bob has spent a lot of time reading blogs that support George C. Marshall Institute's agenda. Perhaps not surprisingly, Bob believes that AGW is not a real phenomenon but rather a conspiracy of some kind – that it is believed only due to certain interested parties manipulating the mass media and the scientific community. Consequently, he also believes he is doing no harm, even though in fact his actions are contributing to AGW and its many negative byproducts in various ways, not least by him openly advocating us to continue business as usual. Clearly, Bob is wrong in his beliefs as the Marshall Institute is essentially only a politically motivated propaganda machine (Oreskes & Conway 2008). The thought experiment then asks: *Is Bob negatively morally responsible for his beliefs, and the subsequent negative actions that significantly follow from his beliefs, concerning AGW?*

Notice that Bob's beliefs, and his subsequent actions that hinge on those beliefs, seem to derive from his constitutive attitudes, dispositions, and traits. For example, as revealed in section 4.1.3.2, Sher mentions an agent's constitutive features to include things like their moral commitments, wits, and biases as defined by cognitive psychologists, for example. Thus, given appropriate background assumptions about Bob, it would seem likely that Sher would consider him responsible – given that Sher acknowledges that Bob's beliefs and subsequent actions relating to AGW have caused and are causing negative consequences.<sup>144</sup>

To ease our examination, let us imagine two instructional heuristics of how one might intuitively answer the question concerning Bob's responsibility:

On the one hand, one might intuitively say that Bob *is responsible*. One might then justify that intuition by saying, for example, that Bob has access to all the information on the Internet and other media, and he *should* use that access to his advantage and carefully revise his beliefs; making him responsible for his continuing omission and his consequent beliefs (and actions). And, taken as a rule, one might argue that the same goes for *anyone* who has similar false beliefs along with basic ability of human reasoning and unutilized access to knowledge – practically leading to universal responsibility for any similar beliefs (and actions) at least in all developed countries with access to

Internet. The implication here is that (almost) *everyone* is *de facto* responsible for their beliefs and how those relate to our current state of important scientific knowledge – regardless of whether they are currently properly aware of that relation themselves. What is also implied is a moral duty for the agents to change their beliefs (and actions) accordingly. From now on, I shall refer to this view as “view 1”.<sup>145</sup>

On the other hand, one might intuitively say that Bob *is not responsible*. One might then justify that intuition by saying, for example, that Bob cannot be responsible, and should not be held responsible, because he lacks relevant knowledge, metacognitive or otherwise, that would allow him to change his beliefs (and actions). And, taken as a rule, one might argue that the same goes for *anyone* who has similar false beliefs. The implication here is that responsibility can only be gained once sufficient awareness of relevant knowledge is gained, thus putting responsibility *solely* on those who have already gained the relevant knowledge. Before gained knowledge has been sufficiently confirmed, judgment of any agent’s moral responsibility should be suspended, for the aim of optimizing effective conveyance of (morally relevant) knowledge. From now on, I shall refer to this view as “view 2”.<sup>146</sup>

View 1 and view 2 represent two conflicting intuitions on whether Bob should be held responsible. In the remainder of this section (6.3, in subsections 6.3.2 & 6.3.3), and in the next section via a pragmatic argument (6.4), I argue for the latter view being better justified and illustrate some problems that I see in the former. While doing so, I hope to illustrate why particular emphasis on (passive) awareness of relevant (meta)knowledge, and emphasis on the engaged perspective, should be encouraged when assigning moral responsibility to someone like Bob.

The arguments I introduce are two epistemically motivated ones. These may be called the first scientist argument, in section 6.3.2, and the epistemic foundations argument, in section 6.3.3. In section 6.4, I also introduce a pragmatically motivated reason that overlaps with a moral reason.

The basic thinking behind the view I’m advocating – that to hold an agent responsible should necessarily require the agent to have gained relevant knowledge – might be summarized in the following thought: once you gain knowledge (e.g., scientific or metacognitive) about something, then and only then can you properly accept responsibility related to that knowledge.<sup>147</sup> I think this amounts to a necessary requirement for a pragmatic view of moral responsibility. To be a bit more specific: I think one should be at least passively *aware of* the relevant knowledge and *understand it*, thus enabling *epistemic acceptance* of the knowledge, for us to hold one responsible (as *psychological acceptance* and/or implementation of the knowledge into behavior may yet require the nudge of responsibility).<sup>148</sup>

### 6.3.2 The first scientist

The first epistemic reason to prefer view 2 over view 1 has to do with a related thought experiment. Let us present a question concerning an imagined moment of knowledge formation in science: *How should responsibility be divided before and after a scientist comes up with vital new knowledge, first in the world?*

To answer the question, let us begin by focusing on the ‘before’. Should we think every individual is *a priori* responsible for something that no one knows? Someone might say that yes, everyone is responsible for the accumulation of knowledge in humanity – and therefore every individual is responsible for discovering new knowledge. And, perhaps by extension, we would all be responsible for the critical false beliefs that are currently held due to us not putting enough effort to discover all the critical knowledge that would be crucial for us to discover. Note that this is in accordance with view 1: as our beliefs play a significant role in our actions and omissions, we *should* all strive to discover critical new knowledge to revise our beliefs where necessary, so as to better inform our actions – regardless whether we have relevant knowledge to do so.<sup>149</sup>

But not only would this assign a goal that *all* humans should strive towards epistemically, more significantly it would merely imply that everyone is *a priori* responsible for accumulating knowledge (to the best of their ability) – which is not the same thing as everyone being or being held *a posteriori* responsible for the implications of a certain piece of knowledge (and for our individual beliefs relating to it). And I think it is reasonable to grant – for both views 1 and 2 – that for us to hold someone morally responsible for something, that something needs to be discoverable and discovered *at least* by someone. Therefore, although a quite similar but more specific discussion could be had on whether everyone should be held responsible for the accumulation of knowledge within humanity, view 1 and view 2 can agree that it would be *impossible* for us to hold anyone responsible for something that *no one* knows.

Now, let us focus on the ‘after’. *Once* the scientist has gained the vital new knowledge, first in the world, should we *now* start to think that people who do not know what the scientist knows are responsible for their beliefs relating to the newly discovered knowledge? I think not, for the knowledge has not yet reached anyone but the scientist. Essentially, everyone else still lives in a world that is crucially similar to them than it was for the scientist at the time before she gained the new knowledge.

However, one might argue that the scientist who has discovered the knowledge *can* now start to hold people responsible in relation to the new knowledge – even though other people cannot start doing so (as they don't yet have the knowledge to be able to). For example, if we assume that there would have been a scientist who first in the world found out about the role of humans in climate change – about the effects of anthropogenic emissions of greenhouse gases (see IPCC 2014, 2) – the scientist could have held all humans responsible right after the discovery.

However, a sensible scientist seems to only assign a certain kind of *causal responsibility*; i.e., that people are responsible in the sense that their actions are *causing* climate change. But, crucially, causal (as well as correlative) responsibility should be distinguished from moral responsibility.<sup>150</sup> And what also needs to be distinguished is *being responsible* for something from *being held responsible* for something (see A. M. Smith 2007; Talbert 2016, 61–63). If we say that the scientist can simply start to *hold* people responsible – even if they really are *causally* responsible – some problems will occur.

*Firstly*, I should further clarify the relevant distinction among the two different forms of responsibility that I'm referring to (causal and moral). I think people can be held *causally responsible* – given that this refers to a purely *descriptive* account of the role people play as part of the natural world – although, due to pragmatic reasons that I will come back to later, we probably *should not* explicitly do so without the agent possessing relevant knowledge *if* the causal responsibility in question implies moral judgment. At the same time, I think people not only *should not* be held but also *cannot be held morally responsible* without them possessing relevant knowledge – given that it implies a *normative* demand that, in my view, requires a conscious entity to have a certain epistemic state, including awareness and understanding of relevant knowledge, for it to function as a proper moral guide for that entity. In other words: I think people *can be* and *can be held* but probably *should not be held* (especially explicitly, publicly) causally responsible without them possessing relevant knowledge when that responsibility implies moral judgment; and they *cannot be*, *cannot be held*, and *should not be held* morally responsible without them possessing relevant knowledge. Thus, overall: I think we *should not* assign neither kind of (explicit) responsibility to people, without them possessing relevant knowledge.<sup>151</sup>

*Secondly*, any rational person who the scientist openly holds responsible – *either causally or morally* – would take that as just subjective ramblings *if* the specifics behind the claimed responsibility, namely the relevant knowledge, isn't spread (including data and justification for the interpretation of the data; or knowledge about why the corresponding scientists, like climate scientists, should be regarded as *reliable testifiers* as a whole). In other words, if the scientist only

assigns responsibility to people, without passing on the relevant knowledge that justifies her assignment of responsibility beforehand – or without making sure someone else has passed on the relevant knowledge beforehand – no rational person is going to take that assignment seriously. This is similar to what happens in the scientific community: knowledge formation comes first, and only then can we make various moral judgments – and accept moral judgments – that utilize the knowledge. Thus, the relevant knowledge *needs* to spread for someone to rationally accept they are justifiably held responsible for something relating to that knowledge. And, consequently, those who already have the knowledge are the only ones who can have justified responsibility relevant to that knowledge. Thus, as knowledge is successfully spread, responsibility is successfully spread – as responsibility can and should be openly assigned to those who *do* have the relevant knowledge.

*Thirdly*, to briefly get back to the pragmatic point concerning causal responsibility: I think that it may very well be pragmatically important to *not openly assign* even causal responsibility without the agent being aware of relevant knowledge, insofar as a particular instance of causal responsibility implies a moral judgment. I think so not only because knowledge needs to spread for someone to rationally accept they are justifiably held causally or morally responsible for something (as discussed above), but also, and more importantly, because the effects of *openly assigning either kind of responsibility without this sort of knowledge requirement might be pragmatically detrimental for our moral goals*. So, in my view, in assigning moral or morally implicative causal responsibility we *should* demand that a knowledge requirement is fulfilled, which requirement includes the target agent having become aware of and having understood the relevant knowledge, thus enabling epistemic acceptance of that knowledge. For example, we should be comfortably sure that Bob is aware of the evidence of AGW and that he understands it sufficiently enough *before* we assign causal or moral responsibility to him (even though he is, in fact, causally responsible). I will address this pragmatic point more specifically in section 6.4.

It seems clear that the above-mentioned problems remain even after the knowledge starts to spread beyond the one scientist: it *should* always be required that the agent we'd like to hold responsible is aware of and has understood the relevant knowledge, enabling him or her to accept it, before we hold him or her responsible. Of course, someone might say that there would be some other criteria to be found after the knowledge formation, or after the knowledge starts to spread, that would sensibly allow us to hold Bob, and almost everyone, responsible. I'm quite unconvinced that there would be other plausible criteria.

Here, those who hold view 1 would need to address how it could be that *sometime* after the first scientist has gained the knowledge, (almost) everyone has gained relevant responsibilities (and,

furthermore, *should be held* responsible). Surely the answer couldn't be merely the increased number of scientists and other people who possess the knowledge – as Bob, for instance, still doesn't. It is not at all clear to me how, for example, Sher would approach this issue. As Sher considers Bob's constitutive psychology the source of his failure, and thus him as responsible, there seems to be some implicit assumption that links the justification to direct moral demands to Bob with Bob still being unaware of the relevant knowledge needed for him to change his beliefs. My best guess would be that the assumption holds the moral demands to be beneficial in eliciting behavioral change in Bob (or, alternatively, merely satisfying some primal urge to hold people accountable as they appear to merit it, and at least without hindering elicitation of behavioral change; see sect. 2.2). However, as seen above, this may be unlikely to work. Further, I will argue in section 6.4 for such an assumption to be *in scientific fact* misguided, and it to be more beneficial to focus our attention to trying to optimally convey the knowledge to Bob.

The thought experiment thus far has tentatively illustrated how knowledge may be needed for justifiably assigning and accepting responsibility. Thus, it may be optimal to adopt view 2, which holds that responsibility is something that spreads at the point when knowledge successfully spreads (or possibly when metacognitive knowledge is acquired; see sect. 6.1.1.5). In other words, when someone – scientist or not – gains new (meta)knowledge, s/he gains relevant responsibility to that knowledge. And as responsibility is gained, similarly the responsibility spreads onwards as the knowledge is successfully spread further.

### 6.3.3 Varieties of epistemic foundations

*The second epistemic reason* to prefer view 2 has to do with the fact that Bob might hold a different epistemological foundation or understanding from that of science; and scientific knowledge itself would therefore be very difficult – if not largely nonsensical – for him to approach and understand.<sup>152</sup> This may be thought in terms of Ludwig Wittgenstein's idea of *hinge commitments*: Bob's epistemological hinge commitments, and the language games that result from them into his subjective perception and reasoning, may be so far off from the epistemological hinge commitments of scientists that even if he were willing to try to revise his beliefs about climate change himself, he might simply be unable (see Pritchard 2018; see also Druckman & McGrath 2019; Garrett & Weeks 2017; Ranalli 2018).<sup>153</sup> And the only thing he would be able to subjectively “understand” might be the unscientific propaganda by the Marshall Institute. Furthermore, his information-seeking skills and the respective information environment might be lacking so much that he would not even be able to find any

counterpoints for the Marshall Institute himself (see Bolin & Hamilton 2018). For example, he might be living in an *epistemic bubble* (Nguyen 2018a, 2018b; see also Pennycook et al. 2018; Nguyen 2018c; sect. 1.1n7). Or, worse yet, he may be living in a strongly fortified *echo chamber* with his social circles systematically and, for him, convincingly discrediting any viewpoints counter to the Marshall Institute narrative (ibid.).<sup>154</sup>

Due to these kinds of possible shortcomings in Bob's epistemological foundation and surroundings, it would be imperative to try to reach and educate Bob – and people like him – about the flaws in his epistemological understanding (and possibly in his moral understanding, metaphysical assumptions, and/or information-seeking skills) (see Garrett & Weeks 2017). And, consequently, it would make little sense to hold him responsible and direct moral demands to him – requiring him to suddenly rebel against growing up and continuing to live in an environment that has shaped his epistemological understanding, constitutive psychology, and relevant abilities and motives and information environments to be such that in their current state they make him simply unable to revise his beliefs by his own means. Of course, reaching and educating Bob is easier said than done (cf. Nguyen 2018, 11), but it is the remedy that is needed, and a good place to start is by approaching Bob in an effort to build *trust*. If anything, trust is only killed via assigning blame if the target is unable to epistemically approach its justification.

To give another, perhaps more concrete example of what I mean when I say Bob might hold a different epistemological foundation from that of science: A fundamentalist Christian might have serious difficulties in approaching even the basic epistemology concerning the science of evolution or cosmology, given that his or her relevant epistemic foundation (i.e., presupposition) has been indoctrinated to lie solely in the literal interpretation of the Bible. Therefore, it seems redundant – to say the least – to direct moral demands to the Christian to adopt evolution (view 1). But it wouldn't seem redundant to put responsibility on those who understand the relevant epistemology concerning evolution and cosmology – and preferably also fundamentalist Christianity – to try to convey that better predicting, testable, and more accurate epistemology to the Christian through education via long term discourse (view 2).

As can be seen, view 2 provides a solution for this problem of epistemological (and moral and metaphysical) foundations as well: it is, already in the first place, the responsibility of those who know better to actively try to educate Bob, and others like him, in the ways required – instead of the misguided, or redundant, requirement by view 1 to just hold Bob *de facto* responsible for his beliefs and actions. As it happens, I would imagine this is largely how Aristotle might have thought about the matter: after all, he does emphasize the role of politics, education, and upbringing in enabling a

propitious societal context where virtue can be cultivated and where people can flourish (see Aristotle & Reeve 2014, X.9; section 2.1n33). Insofar as someone has grown up in and acted in a context where virtue – especially intellectual virtue – has been failed to have been taught or encouraged, and as a result they have either failed to cultivate virtue or have cultivated vice instead, it makes little sense to blame them. Instead, it makes sense to help them by starting to create a context with the desired effect of enabling virtue to arise.

#### 6.4 A Pragmatic Argument: The Case of Jack and Global Warming

Here, I shall get back to the point, already briefly mentioned in section 6.3.2, of how openly assigning responsibility without a knowledge requirement might be pragmatically detrimental for our moral goals.

To clarify what I am about to present, it is helpful to repeat the competing views I initially outlined in section 6.3.1: View 1 holds that assigning responsibility does not require the subject to have any awareness of relevant knowledge by default; leading to some sort of (near-)universal responsibility, exemplified by the case of Bob and his responsibility for his beliefs and actions related to AGW. View 2 holds that assigning responsibility should require the subject to have a certain kind of epistemic state before assigning responsibility, including (passive) awareness of relevant knowledge along with understanding that knowledge, and thus enabling epistemic acceptance of the knowledge.

I would be hesitant to adopt view 1, not only for the epistemic reasons presented in section 6.3, but also due to pragmatic reasons. The following reasoning in support of view 2 puts certain responsibilities on those who say Bob is responsible (view 1), but notably by using their own reasoning for assigning responsibility. It is my intention to thus demonstrate the pragmatic weakness of holding view 1 – by using the reasoning behind the view against itself. My examination first raises a philosophical problem arising from the view 1 line of thinking, namely a paradox and a *reductio ad absurdum* (6.4.1), and secondly, it raises a potentially serious moral problem when considering some relevant research in social psychology (6.4.2). In the end, we will see that those who say Bob is responsible may themselves, by their own reasoning *as well as* by the reasoning via view 2, be responsible to not say that – i.e., be responsible to change their position concerning the epistemic condition. In section 6.4.3, I outline what this would seem to imply for Sher's account and other accounts of the epistemic condition, how the pragmatic view emerging from this examination might be adopted more widely, and I answer some lingering questions.

### 6.4.1 Intuitions and practices

To start off, let me add *Jack* to the discussion, in addition to Bob. Consider the following scenario:

- (1) Jack is responsible for finding out critical knowledge he is unaware of (even if that knowledge goes against the epistemic beliefs and claims in his social circles). [i.e., view 1 is accepted]
- (2) Jack holds view 1 (i.e., Jack thinks (almost) everyone is *de facto* responsible for finding out knowledge about climate change and other critical issues).
- (3) Jack wants to effectively facilitate Bob to change his beliefs about climate change (i.e., Jack wants to help solve AGW).
- (4) Assigning responsibility to Bob can conflict with Jack's moral goal in (3), as
  - (a) assigning science-related responsibility to people whose epistemic understanding is vastly different from the epistemology of science is off-putting to those people, or it is redundant.
  - (b) assigning responsibility to people who do not have the relevant knowledge significantly correlates with passivity of those people, and with passivity of people who *do* have the relevant knowledge.
- (5) Those holding view 1 should not promote their view if they want to have an optimal effect on Bob's and other people's views on climate change. [from 4a^4b]
- (6) Jack *should* stop promoting view 1. That is, by his own reasoning: Jack is *responsible* to stop assigning responsibility to those who genuinely do not know – and possibly regardless of whether Jack is aware of 4a & 4b. [from 1^2^3^5]

Notice how Jack is in a position, in relation to his view 1, where all the people currently holding view 1 may actually be in. They just might not be aware of knowledge in the vain of 4a and 4b; and if evidence to support such knowledge claims has actually been discovered, they would potentially, by their own reasoning, *already* be responsible to abandon their view 1 (or alternatively abandon their moral goals, in which case holding Bob responsible would seem strange). Conversely, those holding view 2 would just be able to say that those who have acquired the relevant knowledge 4a and 4b would be responsible to spread that knowledge to those with the unproductive view 1 of responsibility.

*Firstly*, this situation illustrates to the people who hold view 1 what it *feels like* to be Bob. It illustrates what it feels like to be suddenly held responsible on a significant piece of *alleged* knowledge that you are not properly aware of, having previously managed to miss or misunderstand that piece of knowledge. These sort of vague claims – 4a, 4b – may be comparable to what Bob has ever heard of climate change, or he hasn't heard even an amount comparable to those, or he has only heard false propaganda (that is, having only heard  $\neg 4a$  and  $\neg 4b$ ). This illustrates why Bob might simply be unable to change his views himself, especially considering that his primary source of information is associated with the Marshall Institute. This also illustrates how these kind of vague snippets, which may go against all our current beliefs, merely may or *may not* move people to find out more about the knowledge claims they make (4a, 4b).<sup>155</sup> If I were to stop writing here – and profess view 1 and count on the responsibility of all readers regardless of their epistemic state – I bet many people with view 1, who might be reading this, would not take the necessary steps and try to find out more about the above-mentioned knowledge claims. Instead, they might happily continue with view 1 regardless. This would make their epistemic state regarding view 2 comparable to Bob's epistemic state regarding climate change. Further, in the sense of having related responsibility, through their own reasoning they already *were* comparable to Bob even before this new knowledge, or snippets, had successfully reached them (given that it was available in a reasonably attainable place somewhere, but the attaining of which was blocked by their constitutive psychology).

*Secondly*, the mere possibility of this kind of a situation raises a philosophical problem for those who hold view 1: Jack – i.e., those who hold view 1 – is potentially *always* stuck within his own reasoning. In other words, he may always be responsible to change his views without him ever necessarily realizing it, especially given that view 1 would be widespread (i.e., if everyone would just hold Jack himself *de facto* responsible to find out about the relevant knowledge, without him necessarily ever properly hearing about it). Jack might himself always happen to operate under unfounded assumptions regarding the consequences of the speech acts he performs when assigning responsibility to people like Bob – and might thus always be counterproductive to his own moral cause. Given this, putting emphasis on Jack's responsibility and directing moral demands to him on an issue he isn't even properly aware of is unproductive if not near absurd. Following view 1, the inescapable question follows: How should Jack come to properly know something he doesn't necessarily know, or how should he even come to the knowledge that he should know something he doesn't necessarily know he should know etc.? The most pragmatic answer to the infinite regress and this whole self-defeating paradox and *reductio ad absurdum* seems to be to widely adopt view 2 instead of view 1.

View 2 is not likewise in jeopardy since it promotes responsibility through spreading (meta)knowledge people have already managed to acquire, thus putting emphasis on people always communicating new critical knowledge to more people. And no person is responsible before the knowledge has reached the person. After the relevant knowledge has reached the person (i.e., s/he has become aware of the knowledge and has understood it), s/he can be held responsible for any repeated offences against the knowledge they now have. Therefore, view 2 puts emphasis on educating the public and is also clearly less off-putting and counterproductive in terms of both possibly accurate knowledge claims 4a and 4b. Those who hold view 2, instead of view 1, are less likely to counterproductively judge, stigmatize, or label people regardless of their epistemic state – rather they are directly trying to improve that state by spreading and articulating knowledge, and possible *subsequent* responsibility, via appropriate discourse.

To further underline the more general implications of the aforementioned, they can be presented in simplified formal terms as follows:

- (A) S holds agent X responsible following a blaming strategy V1.
- (B) S holds a moral goal G.
- (C) V1 is demonstrably an unproductive or counterproductive strategy in pursuing G.

Hence

- (D) S should not follow strategy V1.

And further:

- (E) V1 makes it unlikely for a better strategy to be discovered or communicated.
- (F) V2 makes it likely for new strategies to be discovered and communicated.
- (G) V2 is demonstrably a more productive strategy in pursuing G than V1.

Hence

- (H) S should follow strategy V2.

#### 6.4.2 Empirical evidence in the case of Jack

An attentive reader may have anticipated my next point coming: I don't think those who hold view 1 may merely *hypothetically*, for the sake of an argument, operate in counterproductive ways, rather I think they may very well be *actually* doing so. Next, I shall present some scientific support hinting towards that possibility, thus also truly bringing the moral dimension into the discussion.

If there can be found well supported scientific evidence for the knowledge claims 4a and 4b, then it would be empirically supported that holding view 1 is, in scientific fact, unproductive. Specifically, the interest lies in finding out if assigning responsibility to those with different epistemological understanding from that of science (4a) or to the unknowledgeable (4b) is connected with inactivity of the unknowledgeable, the knowledgeable, or both. If support can be found, view 1 may be seen as an immoral view to hold – given that we should want to optimally battle climate change (or whatever comparable moral goals we may have).

Near the end of the previous section 6.4.1, I wrote that those who hold view 1 are more likely to counterproductively judge, stigmatize, or label people regardless of their epistemic state. To further illustrate this point: Explicitly saying, perhaps even just implying, “you are responsible” or “you are to blame” – which often happens via judging, stigmatizing, or labeling people (see also sect. 1.1n3) – might often be taken as a direct blame for one’s behavior and lifestyle, where a moral judgment, accountability, and related demands are cast directly on one’s person and often by extension to one’s social circles, i.e. in-group.<sup>156</sup> Given this, assigning responsibility might often lead to *guilty bias*, provoking self-defensive biases and motivated (moral) reasoning; meaning the individuals held responsible may engage in biased, self-defensive cognitive processes to minimize perceptions of their own and their in-group’s complicity (Markowitz & Shariff 2012; Rothschild & Keefer 2017; Täuber & van Zomeren 2013; see also Popan et al. 2019). Furthermore, as most people probably perceive AGW as *unintentional* (i.e., few people intent global warming to happen), they are prone to judge the harm less harshly than a comparable intentional harm (Markowitz & Shariff 2012). This is all, of course, hindering the ratification of actions required by AGW. But assigning responsibility to an individual is likely to be particularly, if not solely, hindering if we do so without the kind of knowledge requirement described above – that is, before the individual, like Bob or Jack, has enough knowledge to be able to properly justify the claim of them being responsible (Täuber & van Zomeren 2013; see also Täuber et al. 2015). (see 4a and 4b in sect. 6.4.1.)

Beyond the possible guilty bias, some concepts from social psychology that to me seem relevant here include the bystander effect, Dunning–Kruger effect, and equality bias. I see each one of these to be an example that hint towards the possibility that those holding on to view 1 may *actually* be doing so with a significant pragmatic cost to their own moral causes. In the following sections 6.4.2.1–6.4.2.3, I shall go through each of these concepts, respectively, and illustrate how they may support the suspicion of view 1 being counterproductive. In section 6.4.2.4, I will conclude with a brief mention of how knowledge is not all that matters, even though it is our primary focus in this context.

#### 6.4.2.1 *The bystander effect: Dismissing those in distress in larger groups*

*The bystander effect* refers to a phenomenon widely documented in social psychology. It is commonly attributed to have been first discovered in a series of classic studies by social psychologists John M. Darley and Bibb Latané in the late 1960s. The effect seems to be relatively easy to replicate: in many kinds of emergency situations, a person in distress is less likely or slower to get help from bystanders the more of them there are present (e.g., Darley & Latané 1968; Latané & Darley 1970; Latané & Nida 1981; see also P. Fischer et al. 2011).<sup>157</sup> That is, as the number of passive bystanders increases, the likelihood that any one bystander provides help for the victim decreases. Already early on, research suggested some factors that may at least in some situations amplify the inhibition of bystander helping behavior. These include, for example, relative anonymity between bystanders (Latané & Rodin 1969), other bystanders' perceived lack of being alarmed (Darley, Teger, & Lewis 1973; Latané & Darley 1968; Latané & Rodin 1969), and ambiguity of the situation (Clark & Word 1972; Latané & Darley 1968; Solomon, Solomon, & Stone 1978) (for an early ten-year review, see Latané & Nida 1981). A more recent and up-to-date meta-analysis by P. Fischer et al. (2011) has identified some attenuating factors, including the emergency situation being perceived as dangerous (compared with non-dangerous), perpetrators being present (compared with non-present), and costs of intervention being physical (compared with non-physical, especially time or financial). According to the *arousal-cost-reward model*, dangerous emergencies are recognized faster and more clearly as real emergencies, thereby inducing higher levels of physiological arousal and hence more helping behavior, and especially when surrounded by people perceived to be able to provide physical support (P. Fischer et al. 2011).

Although the bystander effect is traditionally associated with emergency situations happening around one's immediate surroundings (as is the focus here as well), similar phenomena have been studied and observed in a wide variety of situations, including in mundane nonemergency settings where group size appears to inhibit individual's behavior (see P. Fischer et al. 2011; Latané & Nida 1981). The effect also provides a potentially useful lens through which to view climate change and responsibility (e.g., Booth 2012; Frantz & Mayer 2009). Notably, the emergency situation concerning climate change does not seem to have any of the attenuating factors present: climate change is usually not perceived to be *immediately* dangerous, there are no *clear* perpetrators, and there are no *clear* physical costs of intervention. Instead, climate change is abstract and distant (Frantz & Mayer 2009; Markowitz & Shariff 2012; Norgaard 2009, 3–5 & 33; Spence, Poorting, & Pidgeon 2012). It may be that a large scale, similar kind of effect plays some part in climate change (policy) inaction.

A simplified classic example of the bystander effect helps to illustrate its connection with responsibility and inaction: A person has collapsed on a busy city street, and a lot of people just walk by without helping. On more rural areas, where there are only a few people around, a collapsed person is likely to get help much more quickly. The less people there are around the more likely a bystander is to help in this kind of a scenario.<sup>158</sup>

In part, the classic example can be seen to demonstrate how people are prone to shift their responsibility to other people they see walking around the situation en masse, leading to *diffusion of responsibility*. For example, they may be thinking that “surely someone will help; I don’t have the time”, “there are so many people around, I don’t want to embarrass myself”, “there must be someone more competent around who will help”, “I don’t want to get involved”, or even “no one is reacting, help is probably not needed”. The problem is that people do not realize most other people have the tendency for the same kind of thought patterns – conscious or unconscious – to rationalize abandoning their own responsibility.<sup>159</sup> To underline: The diffusion is indeed likely to occur the more people there are in a group with a shared bystander responsibility (likely due to people finding it more difficult to defer from the behavior of a larger bystander in-group), and especially when no attenuating factors are present (P. Fischer et al. 2011).<sup>160</sup> It seems likely that the diffusion is also amplified when responsibility is not explicitly assigned, as it may even be eliminated when responsibility is explicitly assigned or elicited (though I am not aware of any experiments that would have controlled group size in relation to committed vs. no-committed bystanders; see Moriarty 1975; Guéguen 2014; Guéguen, Dupré, Georget, & Sénémeaud 2014; Guéguen, Martin, Silone, & Pascual 2016; see also notes 161 & 165). I think all this makes view 1 counterproductive for the cause of addressing climate change and probably for many other causes too, as implying “everyone is responsible” assigns responsibility and a kind of *group pressure* only in a very arbitrary manner, whereas view 2 seeing that “those who acquire relevant knowledge immediately acquire relevant responsibilities” is a much more explicit and targeted assignment of responsibility (thus potentially minimizing the occurrence of diffusion of responsibility). In a sense, the latter appropriately narrows the bystander in-group, thus minimizing the bystander effect. Not to mention it is less assuming of the others’ epistemic state (see sect. 6.3.2 & 6.3.3; see also note 161). (see 4a and 4b in sect. 6.4.1.)

I suspect it is generally accepted that help *should* be provided in these kinds of bystander scenarios. Given a premise that the people walking by accept this normative stance, and think other people accept it as well, it is revealing how differently choosing between view 1 and 2 might affect their approach to the situation. Adopting view 2: every capable human being walking past has a responsibility to help, for example in minimum to make sure an ambulance is called, immediately

*after* they learn that someone needs help (or after they learn that someone appears to need help, calling for confirmatory knowledge acquisition). And a passerby who has adopted view 2 could potentially shift responsibility *only* to people who already *know* that there is someone in distress and who the passerby *knows* to fulfill this knowledge requirement. But since the likely ambiguous situation is happening quickly, surrounded by other busy passerby, they would have hard time telling who else has acquired the relevant knowledge, if anyone, thus making it hard to do anything else but to keep the responsibility themselves (up to the point where help is ensured). But someone with view 1 would be much more prone to shift responsibility, as their view permits the possibility that everyone passing by are *already* (held) responsible to help (and the constitutive psychology of the agents either allows or prevents them from recognizing the situation). Therefore, view 1 would permit a dangerously large harbor to shift responsibility to.<sup>161, 162</sup> (see 4a and 4b.)

The bystander situations are potentially analogous to responsibility in issues like that of Bob and Jack in relation to climate change; i.e., when assigning responsibility to people who don't have proper knowledge of the relevant situation. Essentially, the knowledgeable (e.g., Jack) may be shifting responsibility to the unknowledgeable (e.g., Bob). Furthermore, a knowledgeable person may be less likely to help an unknowledgeable person the more other passive people are around, which could be called the *epistemic bystander effect*.<sup>163</sup> Not only does this seem like the bystander effect in action on a larger scale, fortified via view 1, it also seems to me to be very intuitive on a smaller scale: you are much more likely to have a deep, civil conversation *in private* with your friend or partner, no matter how unknowing you feel they might initially be, than you are to start a similar conversation with a misguided stranger *in public*. In public, resorting to straightforward responsibility without paying attention to the epistemic state of the target seems much more effortless an approach. (see 4b.)

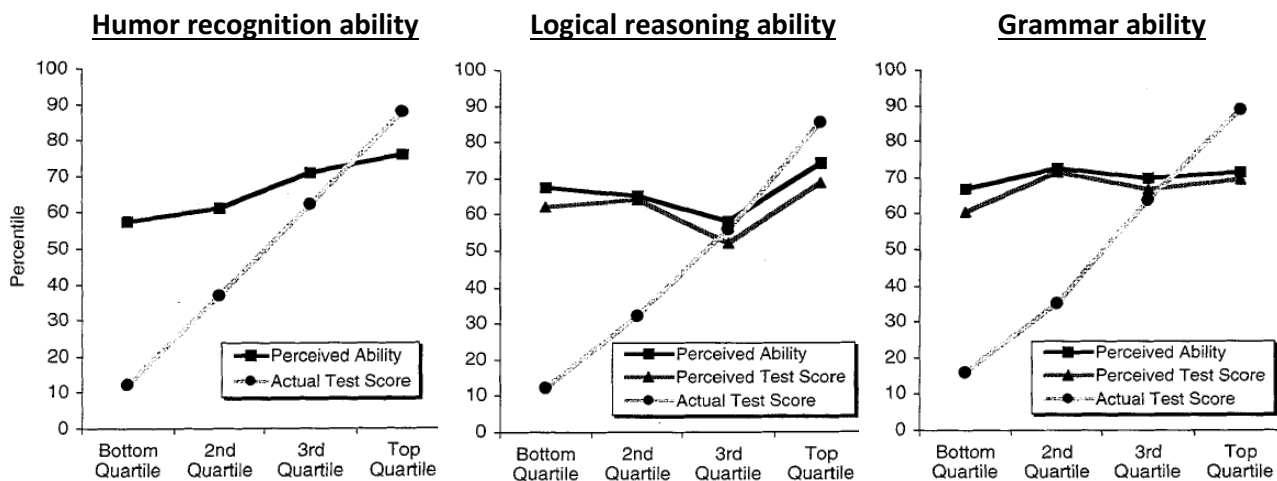
Indeed, in a revealing contrast to the bystander effect being fortified in larger groups, moral punishment seems to increase in presence of an audience (Konishi, Oe, Shimizu, Tanaka, & Ohtsubo 2017; Kurzban, DeScioli, & O'Brien 2007; see also Crockett, Özdemir, & Fehr 2014a). This implies that while we are more likely to dismiss ignorant people in large groups by not providing help for their epistemic state, at the same time – counterproductively, ironically, and almost paradoxically – we are more likely to negatively judge *them* (and leave it at that).<sup>164</sup> (see 4a and 4b.)

The research done on the bystander effect underlines how it doesn't help to shift responsibility to anyone – let alone to someone living in an insufficient epistemic state. Thus, it would be helpful for people to be aware of the bystander effect (and epistemic bystander effect), and the potential tendency to shift responsibility – and we should emphasize that anyone who has come to possess relevant knowledge has immediately acquired relevant responsibilities and *should* start acting like it.

Additionally, this emphasis on view 2 might raise our public self-awareness in situations where we have the relevant knowledge, which in turn might even reverse the potential bystander effect among the knowledgeable (see van Bommel, van Prooijen, Elffers, & van Lange 2012). Furthermore, if passive bystanders can see even one other bystander actively helping the victim, as opposed to just seeing everyone being passive, it can already produce increased helping behavior (P. Fischer et al. 2011). Often, it may just take one ethically well-grounded person to give others a nudge.<sup>165</sup> (see 4a and 4b.)

#### 6.4.2.2 *The Dunning–Kruger effect: You do not know what you do not know*

*The Dunning–Kruger effect* (abbr. DKE) is a phenomenon, or a series of such, first discovered by social psychologists Justin Kruger and David Dunning in 1999. After some minimum threshold amount of information has been acquired on a domain of knowledge/ability/skill, the less competent or knowledgeable a person is on that domain, the more they tend to *overestimate* their competence, both when making absolute estimates and estimates relative to others.<sup>166</sup> In this way, the less competent tend to be unable to metacognitively recognize their lack of competence on a given domain. At the same time, even though the most competent on a domain tend to be more accurate, they also tend to *underestimate* their own level of competence when making relative estimates (see Figure below). When making absolute estimates, the most competent tend to be largely accurate of their own level of knowledge. Furthermore, those who lack knowledge more on some domain are also more unable to properly recognize relevant knowledge (and reliable testifiers) when they encounter it – imagining they *do* have a lot of knowledge themselves. Conversely, as also described in the previous section 6.4.2.1 in another way, the most knowledgeable people tend not to properly acknowledge the lack of knowledge in others. Rather, they are prone to think that if a task is easy for them, then it must be easy for everyone (and hence are prone to underestimate their competence on relative scales; see also false consensus effect: sect. 6.1.3.1n127). These sorts of results have been replicated many times since 1999, on various domains of ability and knowledge, though the precise explanation for DKE is debated.<sup>167, 168</sup> The invisible scope of people’s ignorance is also called *meta-ignorance* (i.e., ignorance of ignorance). (Dunning 2011; Ehrlinger, Johnson, Banner, Dunning, & Kruger 2008; Kruger & Dunning 1999; Sanchez & Dunning 2018; see also Dunning 2014; Dunning, Heath, & Suls 2004.)



**Figure. Composite picture from the first three figures of Kruger & Dunning (1999).**

All participants estimated themselves to be above average (the superiority illusion), but the people in the bottom quartile did the biggest overestimation. In contrast, people in the top quartile underestimated their ability.

(Figure and description adapted from a similar composite used by blogger Neuronicus (2018) with the article copyright 1999 by the American Psychological Association, Inc.)

Noting that most people are far from experts on most domains of knowledge (even though many are experts on some particular area or a couple), this suggests that the most knowledgeable tend not to be well recognized on the collective level, because – due to DKE – the aggregate of people are simply *unable* to recognize them well (Dunning 2011). In other words, the aggregate lacks the necessary (meta)knowledge to accurately evaluate the level of expertise (i.e., to identify and assess the degree of expertise) of someone on a domain where most people are themselves substantially lacking knowledge. For example, as I am currently thoroughly inexperienced and uneducated on the domain of aesthetics in Indonesian literature, say, I largely lack the necessary epistemic tools to evaluate someone claiming to be a master on that domain, and I – likely along with the aggregate of people in Finland, and possibly even in Indonesia – am thus more likely to not accurately recognize them, as compared to experts on that domain. And after acquiring some small amount of information on the topic compared to what is out there – for example, after having travelled in Indonesia or having read a news article on the topic – DKE would suggest I am likely to imagine I know a lot more than I do and imagine I am much more competent than I am (and particularly so when I have only gathered a comparatively miniscule amount of information on the topic, let alone if I have only been exposed to misinformation or disinformation, leading to severe illusion of knowledge).<sup>169</sup> This beginner's inflated view of their ability can translate into overconfidence (Sanchez & Dunning 2018). (see also

Nguyen 2018c for a compassionate epistemological treatise on “cognitive islands”: domains of expertise that are both subtle and isolated in such a way that inexperts are even indirectly unable to evaluate relevant experts; with the domains of morality, aesthetics, and even academics in general being potential examples.)

Furthermore, a related phenomenon of *overclaiming* – that is, claiming knowledge of concepts, events, and people that do not exist – also undermines the perceptions of those who perceive themselves to be knowledgeable. For example, a series of studies found self-perceived expertise within specific domains of knowledge to predict claims of impossible knowledge, particularly claims of familiarity with nonexistent concepts, names, and places on the domain of perceived expertise (Atir, Rosenzweig, & Kruger 2015). There are indications this *may* in some contexts be the case whether or not the perception of expertise is otherwise accurate; meaning that in some contexts even genuine experts may be prone to overclaiming just as well as those whose perception about their own expertise is false (Atir et al. 2015).<sup>170</sup> It appears that we do not only rely on direct examination of our mental contents but also on a *feeling of knowing* (FOK) when we evaluate whether or not we have any knowledge (for a synthesis of two prominent models of FOK, see Koriat & Levy-Sadot 2001; see also Thomas, Lee, & Hughes 2016), and FOK is often only weakly predictive of actual knowledge (Nelson 1984; see also Rozenblit & Keil 2002). Atir et al. (2015) primarily theorize that people’s overclaiming is prompted by FOK, induced by preconceived notions about their expertise that generate top-down inferences about what should be or probably is known – a process that in self-perceived experts can induce a sense of familiarity with terms that sound plausibly real but are not. However, it seems to me to be likely that the problems in our processes of claiming knowledge, and related overconfidence, could be attenuated by cultivating, for example, intellectual humility (Krumrei-Mancuso et al. 2019), analytic thinking (Pennycook & Rand 2019b; sect. 1.1n14), and scientific curiosity (Kahan et al. 2017a; sect. 1.2n25). In a sense, it may be its own form of (meta)ability (or abilities), affected by the Dunning–Kruger effect, to be able to more accurately evaluate limits of one’s knowledge, or the limits of knowledge in general.<sup>171</sup>

Overall, considering both the Dunning–Kruger effect and overclaiming, it seems that few of us are as smart as we think we are. Instead, we seem prone to overevaluate our relative and absolute knowledgeability or (much more rarely) underevaluate our own relative knowledgeability, and respectively being either overconfident or (more rarely) underconfident on a given domain of knowledge. This is emphasized even more if we consider further factors and biases that steer us towards over- or underconfidence; an examination beyond the scope of this thesis (for a useful synthesis, see Moore & Healy 2008).

As Bob – representing similarly unknowledgeable people – is clearly lacking in his knowledge concerning AGW, he is almost certainly unable to recognize that lack of knowledge, as well as likely to be unable to recognize knowledge when he encounters it. Furthermore, Bob is likely to live under a sincere illusion of knowledge, significantly overevaluating his level of knowledge, and possibly overclaiming, while lacking *metacognitive ability* (see Atir et al. 2015; Ehrlinger et al. 2008; Pennycook & Rand 2019c). Therefore, it seems to make little sense to put *de facto* responsibility on Bob himself, rather than to put responsibility on those who already know better to try to effectively communicate to Bob *how* his view is, in scientific fact, lacking – and try to communicate this despite them possibly having a comparatively more doubtful attitude towards their own level of knowledge (see note 167).<sup>172</sup> (see 4a.)

DKE also seems to imply that putting responsibility on those who do not have the relevant knowledge (e.g., Bob in relation to AGW) may be connected with passivity of those who do have the relevant knowledge (e.g., Jack in relation to AGW). To demonstrate why this might be, it helps to emphasize that compared to the knowledgeable, the unknowledgeable people can in some cases be surer of their position (confidently thinking they know more than experts; see note 167) and the more knowledgeable people can be relatively less so, while in addition the epistemic bystander effect may also be part of the equation (6.4.2.1n163). This may manifest as some unknowledgeable people, possibly Bob, being much more vocal in their views *especially* when pressured with responsibility that they see unfounded (due to them unwittingly and massively overevaluating their knowledgeability, and not having the necessary epistemic tools to properly evaluate the epistemic justification for the claimed responsibility). Conversely, the knowledgeable people, at least on some domains of knowledge, may become disproportionately passive: not only due to general challenges of communication, the potential epistemic bystander effect (6.4.2.1n163), and them underevaluating their own relative knowledgeability, but also because they may have more doubts (i.e., less exhibited confidence) about their level of knowledge than the vocal overconfident people have of theirs (see note 167; see also above Figure, logical reasoning ability). In other words, Bob can be much more (over)confident than the knowledgeable (Sanchez & Dunning 2018), and may only be aggravated by claims of responsibility he is unable to epistemically approach. Thus, it seems to me that promoting view 1 may only inadvertently promote the unknowledgeable to more effectively spread their own false views, while, at the same time, not promoting the knowledgeable to effectively spread their views. (see 4a and 4b.)

Conversely, promoting view 2 would strip Bob – and other unknowledgeable people – off responsibility. This would likely make him less vocal as he doesn't feel falsely accused, while putting necessary pressure on those who *do* have the relevant knowledge to become more proactive in *properly* communicating that knowledge.<sup>173</sup> If Bob was to become properly informed from the outside, he might be likely to admit his past errors (Dunning 2011, 274–275), and especially if he is aware that doing so will likely not bring about any negative consequences from the wider community but instead enhance interpersonal impressions (Fetterman, Muscanell, Covarrubias, & Sassenberg 2018). Insofar as he is deeply dogmatic, however, the best we could do would likely be to subtly and compassionately guide him to notice the gaps and inconsistencies in his perceived knowledge and steer him towards a realization of later-to-be-better-informed uncertainty rather than guns blazing attack him with responsibility (Fernbach et al. 2019, 5; see also sect. 6.4.3.1n185).

Overall, DKE can be seen to imply that experts should not hold people responsible for their lack of (meta)knowledge (view 1), rather they should try to educate people; i.e., spread the knowledge along with the only possibly subsequent responsibility (view 2). Ideally, this would also include spreading knowledge about DKE as well as teaching the value of intellectual humility (Krumrei-Mancuso et al. 2019); general analytic and critical thinking skills (Pennycook & Rand 2019b; sect. 1.1n14); and inspiring positive epistemic emotions, like curiosity (Kahan et al. 2017a; sect. 1.2n25). These would likely all help to better live with the effect, both individually and collectively. There are signs that the effect could also be attenuated via teaching people that traits like intelligence are malleable (Ehrlinger et al. 2008; Porter & Schumann 2018). Of course, it couldn't do any harm for us to also seek feedback from those who, for example, have academic credentials to more accurately give feedback on the area of knowledge that concerns us. But before we can properly start to recognize the value in that, it helps for us to first be acquainted with DKE as well as the reliability and limits of academic credentials – all necessarily requiring effective communication from the knowledgeable to the unknowledgeable. (see 4a and 4b.)

#### 6.4.2.3 *Equality bias: All in-group opinions matter as an aggregate*

A group level relative of the Dunning–Kruger effect is *equality bias*. This is a rather novel concept that describes a way in which collective decision-making within and across cultures can be impaired (as found in samples from Denmark, Iran, and China): people seem prone to give equal weight to views of other members in a small group, regardless of the members' relative competence or reliability. This prevents optimum decision-making from taking place as people seem to have the

tendency to downplay their own competence and upsell their lack of competence for reasons relating to diffusion of responsibility (partly via the Dunning–Kruger effect), and/or aversion to social exclusion (likely due to evolutionary incentives regarding group cohesion). There seems to be some significant amount of cognitive and group pressure to include all perceived in-group views as equal, so to not hurt any feelings, likely making the collective decision a sort of average of everyone instead of one that would most likely benefit everyone.<sup>174</sup> (Mahmoodi et al. 2015; see also Bahrami 2018.)

When we notice that there can further be polarized groups or echo chambers within and across societies that each have their own *distinct* internal diffusion of responsibility and/or aversion to social exclusion – grounded in group-specifically perceived principles or moral values – with possibly also group-specific preference falsification taking place (see sect. 1.1n5; sect. 6.4.2.4n176), this equation can lead to an essentially self-enforcing circle of polarization (see sect. 1.1). The circle would continue for as long as no one (or no subgroup of people) breaks through the diffusion and/or aversion to social exclusion prevailing within the groups.

If we take view 2 to have been widely adopted, it would, much like in the case of DKE, seem to put pressure on those who *do know* to not succumb to such group pressure as they alone have the responsibility that the knowledge they possess brings: namely, to spread that knowledge successfully and hold to their guns, so to speak; tweaking their arguments and communicative strategies, and, of course, to their best ability consistently making sure they do have the relevant knowledge and are not gravely mistaken (partly via being informed by those who are more knowledgeable on areas where they lack). Whereas with view 1, everyone would be responsible right from the start of the decision-making process – thus, again, more likely encouraging Bob and other unknowledgeable people to voice their false claims of knowledge, and who could capitalize on the pressure of social inclusion that view 1 does not address as efficiently as view 2. (see 4a and 4b.)

#### 6.4.2.4 *Beyond knowledge*

The biases and effects discussed above indicate why 4a and 4b, that were presented in section 6.4.1, may be accurate – that is, why it may empirically be the case that it would be counterproductive for Jack’s moral goals if he were to hold Bob, and people like him, responsible prior successful conveyance of relevant (meta)knowledge. Furthermore, they illustrate why it would make little sense to hold Bob *de facto* responsible in any case, as there are sociopsychological and other background causes that have their incessant hold on Bob (see also Caruso 2018, sect. 2.5). Those background causes can successfully be addressed only by interventions that take note of, and break through, the

relevant phenomena. And those who already know better – about climate science, and relevant epistemology, and ideally about relevant sociopsychological phenomena – are in a key position. They know the relevant facts and are therefore responsible to try to intervene the lives of those who are the unfortunate prisoners of a limited epistemic state. In fact, they are the only ones who *can* have that particular responsibility as they are the only ones who possess knowledge needed to act accordingly.

It should be underlined that this empirical list is only a tentative review of the evidence hinting at 4a and 4b. The most crucial thing to note is that the pragmatic question about the effects of assigning responsibility with or without a knowledge requirement (or with or without awareness), in specific cases, is a question that can be studied empirically. Most likely, there exists more relevant evidence and related cognitive biases that are not noted here, and further research should be encouraged.<sup>175</sup> Of course, it is also possible that some of the presented empirical considerations are in some way misguided, and/or I might be operating under some distorting amount of confirmation bias. If that was revealed to be the case, then the pragmatic view should be iterated accordingly: in a way that our habits of assigning responsibility could be steered towards the most pragmatic view in terms of fulfilling the aims of our usual moral demands (i.e., to fulfil our moral goals via eliciting cognitive change in the target agent), while continuing to be aware of our fallibility, and optimizing the flow of knowledge in society. I hope that what I have presented has provided at least a decent initial case for considering view 2 instead of view 1. Of course, even if there was no empirical evidence, the arguments in sections 6.3 and 6.4.1 would still be intact.

To conclude this empirical examination, it should be noted that there exists a considerable amount of evidence to support the notion that merely communicating knowledge is *not* all that matters in successfully inducing epistemic and behavioral change in an audience or an interlocutor. For example, also the way knowledge and the issue in general is framed matters (see, e.g., Markowitz & Shariff 2012; Nisbet 2009; Norgaard 2009, 39–42). Especially relevant for a discussion about morality is how *moral reframing* can matter: we tend to frame matters that are important to us only in terms of the values of our moral tribe and the context it perceives, when in fact it would be more persuasive to frame them in terms of values the audience or interlocutor finds important (see Feinberg & Willer 2013, 2015; Hornsey & Fielding 2017; Voelkel & Feinberg 2017; Wolsko et al. 2016; see also Albarracín & Shavitt 2018; Kahan et al. 2017a; Täuber & van Zomeren 2013).<sup>176</sup> The motivation for moral reframing is provided via the large body of research suggesting that what is actually more important than being sufficiently informed is the social and political – and thus moral – group one identifies with (see, e.g., Kahan et al. 2012; Cohen 2003; Hornsey et al. 2016; Pew Research Center

2017d), even though knowledge can also make a difference (Guy et al. 2014; Ranney & Clark 2016; Shi et al. 2016; Weisberg et al. 2018; see also Milfont et al. 2017; section 1.2). People often utilize their acquired knowledge and skills of reasoning to rationalize and support their pre-existing worldview, and in-group ideology, rather than utilize them to genuinely better understand the world, let alone make informed decisions (Markowitz & Shariff 2012, 244; see also sect. 1.1 & 1.2; Haidt 2012). Therefore, it should additionally be emphasized that efficient framing and communicators targeted to specific groups would also be a beneficial approach to adopt when designing communication strategies for encouraging belief and action change in others.

Falling within this general topic of knowledge being only one of several factors in successful communication, there are many nuances that could be discussed further (see sect. 7.3). However, I am mentioning this only briefly, merely as a reminder that conveying knowledge is *not* all that matters in the pursuit of pragmatic effectiveness in seeking to fulfill our moral goals. Knowledge is merely the main topic at hand in this context.<sup>177</sup>

#### 6.4.3 How to adopt the pragmatic view, and implications for other accounts

Having completed the pragmatic examination, in this section I focus on outlining specifically what it implies for other accounts of the epistemic condition, and how it's adoption might be encouraged. In section 6.4.3.1, I outline the implications for Sher's account, and provide some useful exploration to how the pragmatic might be adopted more widely. In 6.4.3.2, I propose and answer some lingering questions about view 2, i.e. *the pragmatic view*. And finally, in 6.4.3.3, I present a summary of some key philosophical differences between view 2 and Sher's account, and illustrate how view 2 might (or might not) conform with other accounts that have been discussed in this thesis.

##### 6.4.3.1 Implications for Sher and challenges of adoption

In section 6.3, I presented an epistemically motivated thought experiment, accompanied with epistemic arguments, to illustrate why view 2 would seem better justified than view 1. Further, in sections 6.4.1–6.4.2, I presented a philosophical pragmatic argument in support of view 2 as well as evidence from social psychology to support view 1 being counterproductive in terms of best fulfilling our moral goals. Both of these can be viewed as arguments to undermine Sher's FEC clause 2 (see sect. 4.1 or 4.2), insofar as Bob and Jack are considered to fall under that clause, which was represented by view 1 judgments of them. As mentioned in 6.3.1 (n144), it isn't clear to me why Bob – and by extension Jack – would not fall under the clause. The case of Bob and Jack acting badly due

to ignorance of the evidence seems to be especially comparable to Sher's example cases *Bad Policy* and *Bad Weather* (see sect. 3.3.2; cf. 6.1.1).<sup>178</sup> Just as Sylvain continues to act unfairly due to his ignorance (or unawareness) of the deficiencies of emotional empathy (see 6.1.1.3), and Amerika preparing and killing for a revolution due to his ignorance (or unawareness) of the deficiencies of his ideology (6.1.1.3), Bob and Jack seem to be acting in comparable ignorance (or "lack of moral insight or imagination"). Thus, it seems the arguments presented against view 1 apply against Sher's clause 2 as well.<sup>179</sup> In terms of the question presented at the end of section 3.3.2, of whether we should rather question our intuitions or to question the role of awareness in responsibility, it now seems that we should rather question our intuitions, if we want to take our moral aims seriously.<sup>180</sup>

However, given that Sher's account or some other non-pragmatic account is, by and wide, a representation of the intuitions of the majority, it may be asked if view 2 is too demanding of people.<sup>181</sup> This seems to be a legitimate concern: how realistic is it to ask people to change their intuitions regarding responsibility and agent evaluation? I am not sure, but the more people could adopt view 2 the better the outcome would seem to be. This doesn't mean everyone needs to change their intuitions: it already helps that people who are disposed to change theirs keep working at it, and people who already possess different intuitions to Sher's are encouraged to acknowledge how their view might be beneficial to tweak and embrace.

In the process of wider change and adoption, for example *metacognitive training* (e.g., Callender et al. 2016; Stokes 2012), *perspective taking training* (e.g., Hooper et al. 2015), *compassion interventions* (e.g., Klimecki et al. 2013, 2014), as well as some (further) forms of *mindfulness or meditation* (e.g., Condon, Desbordes, Miller, & DeSteno 2013; Hopthrow et al. 2016; Lim, Condon, & DeSteno 2015; Wang, Geng, Schultz, & Zhou 2019) may be seen to play beneficial roles, especially if properly implemented to our educational institutions.<sup>182</sup> If ethical reasons are not enough, this implementation could be motivated by the mere learning benefits these methods are likely to bring, while their effects on our moral intuitions could be considered a welcome byproduct.<sup>183</sup> Raising awareness of the importance to effectively communicate knowledge would likely motivate us to not only follow through with the subsequent responsibility, and accordingly practice more pragmatic agent evaluation, but also better pre-emptively fact-check our own beliefs; as we are then responsible to others, not just to our biased selves.<sup>184</sup> Simultaneously, we ought to strive to become better self-aware of our many biases, and systematically meditate on them (e.g., Hopthrow et al. 2016). At the very least, it seems plausible that a considerable number of people could learn new heuristics to utilize in their moral judgments: if we can reflect our judging detached perspective from a further critical

detached perspective, we may come to engage with the world in a more fruitful way (see also Flanagan 2017, 157–216; sect. 6.6).

While encouraging critical evaluation of our detached perspective, the pragmatic view can also be seen to encourage us to reflect on the importance of the engaged perspective – after all, the perspectives seem to be significantly interlinked (see sect. 6.2.1). Particularly, in the thought experiment, Bob and Jack are victims of their respective limited epistemic states, which are essentially cognitive states that concern their engaged perspectives in relation to their beliefs and actions. And the only thing that can free them from their limited states are outside agents who notice those limits within Bob and Jack’s perspectives, manifested via all the possible beliefs and acts that imply deep unawareness of some crucial knowledge (in Bob’s case unawareness concerning, for example, media literacy, science literacy, and AGW; and in Jack’s case unawareness concerning effective communication to people like Bob). Only these outside agents, who know better, can set out to help the misguided Bob and Jack by addressing their mistaken (meta)knowledge and consequent behavior. And they should aim to do so in the most pragmatic ways they know how to convey information and illicit cognitive change. In the process, they are also the only ones who can enable Bob and Jack to rationally accept responsibility to change their actions, after successfully passing on the required knowledge (6.3.2 & 6.3.3). Of course, some misguided people may find and fix their errors all by themselves, but we should never blindly count on it, because we do not know how deep a mistake or an error in another’s epistemic state goes (6.3 & 6.4).

It should be underlined still that if due to our communicative efforts someone misguided – like Bob or Jack – comes to subsequently ‘accept (taking) responsibility’ relevant to the knowledge they previously lacked, as is the goal, this does not mean they ‘accept being responsible’ nor that they ‘were responsible’ for their relevant beliefs or acts that were done *prior* successful belief revision (6.2.2). That is, when Bob or Jack accept responsibility it does not follow that they *were* responsible. Indeed, it seems they were not responsible if they were acting in ignorance of some relevant (meta)knowledge, insofar as they display regret for acting how they did after they successfully gain the (meta)knowledge (6.1.1.4). When they accept responsibility – which seems to often happen near the first point of feeling other-regarding sorrow and regret connected to the gained knowledge – what they are accepting is that they previously lacked important (meta)knowledge that they are henceforth aware of and willing to incorporate into their future decision-making processes lest they *otherwise* become responsible for related future decisions, and liable to be held responsible, for now they possess the requisite (meta)knowledge. For our moral goals, this appears to be all that matters.

One additional way to think about these issues, and one that also illuminates how we might want to think about effective persuasion, is illustrated by social psychologist Jonathan Haidt's rider-and-elephant metaphor, briefly mentioned at the end of section 4.1.4, in note 81 (see Haidt 2012, 52–83). Essentially, one ought to try to appeal to and affect the elephant (*a priori* automatic processes), not the rider (*post hoc* reasoning). When one encounters a rider, and asserts responsibility, the rider will likely continue rationalizing the elephant's movement who may be reacting aggressively to being challenged on a basis it is not properly aware of (e.g., when holding Bob or Jack responsible before properly communicating the relevant knowledge to them) – in which case responsibility as an action-guiding tool can quickly fail its intent. So how to effectively appeal to the habits of the elephant instead? Via the right kind of communication that compassionately thinks before it judges, while speaking to the elephant when communicating the knowledge. In addition to presenting evidence – and even more importantly – this might include, for example, moral reframing, analogous thought experiments, Socratic questioning<sup>185</sup>, principle of charity, assumption of sincerity, making sure the matter is not merely of terminological or definitional in nature, finding ways to replace any necessary emotional needs the false belief might serve<sup>186</sup>, and generally forming a *positive* and *trusting* relationship with the rider and elephant rather than just letting your own reactionary elephant take the rein. In effect: try to get the target elephant to persuade itself. Conveniently, at the same time, we would seem to open our own elephant to better accept its own fallibility, and thus open ourselves to new viewpoints and arguments.<sup>187</sup>

Following this note from Haidt, we may find one answer to the question proposed at the end of section 4.1.4 (see also sect. 4.1.3.2–4.1.3.5; 5.2; Waller 2014, 639–641): why should merely the part of us who is aware be responsible, when the thought processes that we are aware of rely on processes we are not aware of? One answer could be: Because it seems misguided to hold the part of us who is unaware directly responsible via the part who is aware, when the actual culprit in (faux-)control is much deeper inside us (even deeper than any information we have acquired but are unaware of). And, more importantly, because most often the actual culprit seems not to be efficiently corrected via assignments of responsibility. Instead, most often it seems best corrected either via very careful and considerate discussions with the agent who represents it, by those who know better, or via new (meta)knowledge being accumulated in the agent experientially without such communication. In a sense, Sher is arguing that we hold the elephant and rider responsible and direct blame at them, but that is exactly what we ought not do: it encourages reactive dogmatic response (in those who do not possess knowledge relevant to the responsibility), instead of encouraging openness and building a trusting dialogue.

#### 6.4.3.2 Questions about the pragmatic view

In this section, I pose and answer some lingering questions that the pragmatic account has raised. These should further clarify the view.

Question 1: If knowledge should come prior to responsibility, how can those who possess knowledge about AGW or social psychology, for example, be responsible to suspend their likely intuitive judgment on Bob and Jack's moral responsibility and instead only try to effectively communicate with them?

It is true that we cannot hold anyone responsible to suspend their judgment on the agents' responsibility by default. People would first need to become aware of, and properly understand and internalize the kind of philosophical theses and their justifications as presented here. They would become responsible when they realize they are the only ones who can have responsibility relevant to the (meta)knowledge and justifications they possess, and that others do not necessarily possess that knowledge. This ought to give them justification and motivation to try to most effectively communicate with others, which, if I am correct, would imply adopting view 2.

Question 2: What about research that seems to conflict with the empirical evidence presented in section 6.4.2? For example, research about "guilt appeals" (see, e.g., Graton, Ric, & Gonzalez 2017; Xu & Guo 2017)?

The question assumes that guilt appeals are comparable to assignments of responsibility – which I am not sure if they are. In many cases, they would seem to be at most some specific form of *implicit* assignment of responsibility – which mere communication of information can also be, though ideally in a more neutral way still. In any case, it seems to me that the value of guilt appeals is also optimized via view 2. That is, it seems most effective to make appeals to those who have come to possess knowledge relevant to the appeal, and otherwise the appeal may not reach its aim, or it may even backfire (see Wonneberger 2017).

Still, *insofar* as guilt appeals, or, for example, mere group pressure, can encourage people to change their transgressive behavior, at least in the case of AGW they should be encouraged, especially via intergovernmental regulations and transnational treaties. However, mere guilt appeal or group pressure having moved someone to conform does not mean they themselves have sufficient justification based on understanding the evidence (or based on understanding why some people can be considered reliable testifiers to that evidence). And, of course, guilt appeals as well as group pressure can

themselves be often misguided in relation to proper evidence. Thus, to optimally eliminate the chance of people at some point rebelling (if they are the uninformed target of the appeal) and to optimally address the possibility that unsubstantiated evidence is utilized (by the ones making the appeal), as well as to make social deliberation more effective in general, it would seem most reasonable to encourage adaptation of view 2 – even if guilt appeals are at the same time utilized in some carefully considered contexts.

Question 3: Couldn't we just deal responsibility however we like, insofar as we then commit to defend the possible further assertions (epistemic or otherwise) that lay behind our assignment of responsibility? This way we would spread the relevant knowledge and justifications, but only after first assigning responsibility.

Granted, it would seem most important that the knowledge gets to spread *at some point*. However, there are some problems with this suggestion. One problem seems to be revealed when spectating the discourse space on social media (see sect. 1.1; 6.4.2.1n164): many people largely just assign responsibility, and then leave it at that (and this often happens rudely and on misguided premises about what happened, I might add). Of course, this is no worse a problem than the one view 2 seems to face: majority seems not to have adopted this way of thinking. But there is also a bigger problem with the suggestion: the rationalizing, defensive psychological effect seems to be immediate when assigning what the agent's intuitive elephant judges to be *unfounded* responsibility, and thus the damage is already done even if immediately afterwards trying to communicate relevant knowledge to the rationalizing rider. Thus, view 2 would seem like the better option, encouraging us to suspend judgment on the agent's responsibility, and instead focus on proactively communicating the relevant knowledge that would give sufficient reasons for the agent to change their beliefs (and subsequently take responsibility).

Question 4: What would this entail for democratic decision making, laws, or international treaties? Would it mean, for example, that all kinds of collective responsibility would be misguided as not everyone knows all the relevant laws and treaties?

It seems to me that the commonly understood notions of responsibility in terms of obeying the law or following international treaties would be little to no affected. Legal responsibility and democratic attendance are not the same as moral responsibility. This is perhaps clarified, also in any further sense of collective responsibility, by noting that democratic institutions and the rule of law necessarily require only the executive institutions to be informed in some organized manner (even

though it is advisable for the educational and media institutions to communicate any relevant knowledge to the wider public). If someone is not aware of some law, they can still be legally responsible, and they can still be held legally responsible if the executive institution finds out about them performing illegal actions.

Strictly speaking, this is separate from the interpersonal moral responsibility in question in the overall discussion about the epistemic condition. Any collective responsibility related to citizens of democratic governments seem to arise via citizens being engaged in citizenship – albeit usually by force of birth. Thus, insofar as Bob is living in a democracy, he is presumably committed to following, for example, carbon taxes, should they be legislated, despite him not believing in AGW. To follow any democratic laws is Bob's first-order duty, while communicating about any law that he sees misguided is a second-order duty. The collective responsibility arising in a democracy seems to significantly concern the willingness to follow the laws that are in effect, excluding civil disobedience, while also enjoying from the privilege of being free to campaign to change any laws one sees to be misguided. In other words, the privileges of democracy entails commitment to any democratic decisions made, and a freedom to try to affect future decisions the best one sees fit.

That being said, if Bob does not know why democracy is a better form of government compared to any other alternatives (assuming it is), or what a proper citizenship entails, or why a given law is or is not good for the society, this should be communicated to him in a pragmatic manner also, rather than hold him *de facto* morally responsible for his (potential) ignorance and actions that accompany and signify that ignorance. *Ideally*, this communication would also be done prior to any relevant law changes, though that seems an unfortunately tall order for all the Bobs and Jacks and newborns in the world and for all possible changes in law (especially insofar as view 1 is widely adopted). Still, it is something to strive for. In any case, even after any changes in law, it seems most pragmatic to not hold Bob morally responsible insofar as he genuinely lacks relevant knowledge. Instead, we ought to try to communicate that knowledge to him the best we know how, even if it is amidst him misguidedly campaigning or acting against some new (or old) law.

Question 5: Still, isn't the pragmatic view simply asking too much of people if they ought to suspend their judgment about moral responsibility until relevant knowledge would have been communicated to the target agent(s)?

To be clear, it is not so much that the pragmatic view asks much anything of people *per se*, it is merely asking us to pay more attention to the knowledge requirement and to what could be done to

pragmatically account for it and to implement that account into our communication. Practically, what is being asked is for the pragmatic view to be well noted and implemented into our educational institutions, via implementation of relevant philosophy, psychology, and critical thinking into the curricula. And, further, what is being asked is for the pragmatic view to be noted in the new media landscape and amplified via educational communication where possible. These are all very well achievable, practical steps that can be taken to mitigate the effects of the current times of polarization (see sect. 1.1 & 1.2). Additionally, anyone who finds these arguments appealing can and ought to make a personal project out of reflecting the aspects relevant to one's own thinking and behavior, and how one could become more productive and prosocial in communicating across groups.

Question 6: The pragmatic view advocates for cultivating what seem to be minority intuitions relating to agent evaluation and communication, and thus the focus is on cultivating the (meta)cognitive processes of the evaluator. Why not instead advocate for cultivating some minority intuitions relating to knowledge acquisition, thus focusing on the target agent(s)?

Because we do not know what we do not know (see sections 6.4.1 & 6.4.2).<sup>188</sup> Of course, we should *also* advocate for people to acquire important knowledge with the best means reasonably attainable to them and for them to cultivate relevant habits. However, no matter how perfectly we may have managed to cultivate those habits (and few have come very far), we may still never know what we do not know unless someone effectively enlightens us. The responsibility thus always necessarily returns to the knowledgeable to either directly or indirectly help the epistemic states of others (given that the knowledgeable have either made the necessary realizations themselves or have read and understood these sort of philosophical theses).

Moreover, for people to cultivate effective information-seeking habits, relevant (meta)knowledge needs to be spread *a priori*, before we can be sure they even necessarily know what effective information-seeking habits would look like and why they would be important, and thus before they can rationally accept any related responsibility. Insofar as people are aware of what those habits would look like and why and how they would be important to cultivate, we may consider them responsible to cultivate them (within some reasonable boundaries of what can be expected of them as the temporally limited, fumbling, and fallible humans that we appear to be).

#### 6.4.3.3 *Philosophical similarities and differences between the pragmatic view and other accounts*

Here, I first outline some key differences between the pragmatic view discussed and Sher's account. The categorizations are rough ideal types but provide some illustrations to elucidate the differences between the two approaches. Further, I illustrate how the pragmatic view could conform with the other accounts discussed: namely, the (unqualified) searchlight view (see sect. 3.1 & 5.1), Zimmerman's qualified searchlight view (see sect. 5.1 & 5.4.3), and Smith's attributionism (i.e. answerability; see sect. 5.2 & 5.4.4.4).

It was already established that whereas Sher's view is a *merit-based* view regarding responsibility, mine is a *consequentialist* one (sect. 2.2 & 6.1.3). It may further be described that Sher's is an *internalist* account of the source of responsibility, as it pays particular attention to the agent's constitutive psychology. In contrast, the pragmatic account would seem to be more so *externalist*, while paying particular attention to the people around the agent, and the epistemic state of the agent that those people – and often *only* those people – can affect. Sher's account may also be described as *individualistic*, emphasizing not only the individual's constitutive psychology but also his own intuitions, focusing on the target agent being evaluated (from the detached perspective).<sup>189</sup> The pragmatic account would seem more *collectivist*, emphasizing the examination of how we could best function as a collective of fallible individuals, and more so focusing on how agents could best help others (while also more thoroughly noticing the engaged perspective). Both accounts can be read to be normative, however Sher's account seems to be aimed at being widely descriptive (see Sher 2009, 153; Zimmerman 2009, 254). A final distinction could be that whereas Sher focuses on the psychological causal constituents of the agents' actions as the culprits behind their wrongdoing, the pragmatic view tends to see the agents more so victims of that causal structure, especially insofar as that structure lacks awareness of (meta)knowledge that could, if fed into their cognitive system, redirect (and/or metacognitively reinforce) their thinking and behavior in ways that they would themselves judge to be beneficial *a posteriori*.

It seems quite clear that the pragmatic account does not match well with Sher's account. However, *to some degree*, it does seem to be compatible with not only the searchlight view, but also with Zimmerman's qualified searchlight view, and possibly Smith's attributionism. The compatibility of the pragmatic view with other accounts seems to hinge, at least initially, on the question of how they would approach Bob's responsibility (see sect. 6.3.1).

As established, Sher's account would appear to hold Bob responsible. However, the searchlight view would not: Bob is not even passively aware of the knowledge of AGW, and hence is not

responsible. Zimmerman's qualified searchlight view would also most likely not hold Bob responsible, though with a significant caveat: Bob would be responsible only if his ignorance can be traced back to a belief or act he was aware to be overall morally wrong. Given Bob's background, it seems unlikely that there would be such a belief or act, or at least it seems difficult if not impossible to discover in practice.<sup>190</sup>

Smith's attributionism seems to give mixed and hard to interpret signals: On the one hand, Bob's judgment about AGW reflects Bob's sincere evaluative judgment about the *state of evidence*, from which evaluation his morally lamentable actions derive from, and hence he might be responsible insofar as that judgment may be interpreted to display lack of moral concern. But, the judgment seems to be sincere, with no ill intent, and it can be added that it could be carefully formulated, albeit with the standards of a misguided epistemic foundation (cf. 6.3.3). Also, Bob's judgment about the state of evidence does *not* reflect any judgment of his about the bad epistemology he is utilizing being better than the proper scientific epistemology, as he is not properly knowledgeable of the latter (if even of the first). In other words, Bob's bad epistemic thinking does not reflect his moral judgment as, presumably, he would rather believe things that are most likely true but just happens to unwittingly utilize bad epistemic and consequently moral thinking.<sup>191</sup> Insofar as this sort of interpretation can be made where Bob is not responsible, attributionism might seem potentially compatible with the pragmatic view, to some degree. (see also Robichaud 2016, for how difficult it appears to be to interpret whether agents like Bob are responsible under (at least some) quality-of-will accounts, Smith's account being one variation).<sup>192</sup>

However, where Smith's attributionism may fail from the pragmatic point of view might be it potentially encouraging us to assign responsibility prior communicating relevant (meta)knowledge. To reiterate the above-mentioned: It seems that evaluative (moral) judgments of agents, that they would be willing to defend, may in some cases rely on lack of (meta)knowledge. Insofar as attributionism encourages us to not appropriately communicate relevant (meta)knowledge to the agent in such situations, prior responsibility, it may be incompatible with the pragmatic view. Still, this concern might be circumvented by interpreting that false knowledge behind an agent's defense of their evaluations would necessarily imply that the agent's evaluations do not represent a *genuine* evaluative judgment on their part because they lack (meta)knowledge about accurate knowledge or epistemology, or, perhaps, moral understanding (cf. 6.3.3). Curiously, though, if an agent does not lack any relevant (meta)knowledge, it is not clear how often, if ever, they would be willing to defend ill acts. Nor is it clear how well we could recognize and heal lack of knowledge guiding an agent's judgment if our account of responsibility does not explicitly encourage communication of

(meta)knowledge prior responsibility, thus focusing on helping the agent's epistemic state. Thus, pragmatic applicability of also Smith's attributionism seems to be in some considerable doubt, and hence her view may not be attractive from a consequentialist perspective.

Still – if not only for the reason to illuminate the pragmatic view a bit further – it may be useful to also consider that Smith's attributionism is a non-volitional account of responsibility: the control condition is rejected (A. M. Smith 2008; see also McKenna 2008). The searchlight view and Zimmerman, on the other hand, do not reject it. Smith views that an agent can be responsible insofar as the agent's actions reflect their evaluative judgments that they could be asked to defend, and it is of no consequence whether or not those actions are in their control. If the agent is willing to defend immoral actions, then they are responsible. So, how does the pragmatic view approach voluntariness and control?

For the pragmatic view, what primarily matters is that morally relevant (meta)knowledge is successfully exchanged or otherwise gained within the collective of individuals, and that after an individual has gained the knowledge, they can be held responsible insofar as they are deliberately acting against their gained knowledge. Furthermore, the pragmatic view sees that it is not always clear whether a given action is (normatively) immoral, which also might require dialogue within the community, prior responsibility (see section 6.1.1.3, the cases of Ryland and Amerika). In addition to these general descriptions, it would seem that the pragmatic view understands control roughly along the lines of a person being able to perform the required redeeming actions without compulsion, be it external (e.g., someone or something physically preventing them) or internal (e.g., an involuntary glitch or a mental illness or brain tumor overtaking their cognitive system). Insofar as the agent has gained the relevant knowledge and *yet* continues to act against that knowledge – while *not* being involuntarily prevented via compulsion – they are and can be held responsible.

Thus far, from the views that have been properly enough examined in this thesis, it seems that the pragmatic view is potentially to some degree compatible with the searchlight view and Zimmerman's qualified searchlight view. At the same time, its compatibility with Smith's attributionism seem to be in some doubt, and it would seem to be not compatible with Sher's FEC. However, none of the positions examined puts emphasis on communication, which puts their consequentialist value to some doubt, including the two variants of the searchlight view. Furthermore, they do not seem fully compatible with each other, even if they might agree in the case of Bob (see sect. 5.4). It seems to me that to properly understand the place of the pragmatic view among the dominant views of the epistemic condition, a clarifying step back is in order.

## 6.5 A Clarifying Step Back: A More Distant Vantage Point

In this section, I take a step back to examine the discussion, and outline some critical aspects of its general characteristics. In section 6.5.1, I introduce two perspectives within the field to how we should think about our varied, seemingly incompatible attempts to account for the epistemic condition: it may not be that there is only one epistemic condition, as has been implicitly assumed thus far, but many. In section 6.5.2, I illustrate how this implicates that the pragmatic view is distinct from many of the prevailingly descriptive views examined, via its distinctly normative nature.

### 6.5.1 Different types of moral responsibility?

The searchlight view, Sher's FEC, Zimmerman's qualified searchlight view, Smith's attributionism, the pragmatic view, and other positions seem to be *prima facie* well comparable and appraisable against one another. However, as for example Talbert (2011), Tognazzini (2010), and Zimmerman (2009, 2017) have noted, there may be several different modes or forms or types of moral responsibility or evaluability in play between different accounts of the epistemic condition, meaning that there might be a lot of talking past one another going on. Notice that already the distinction between responsibility as accountability on the one hand, and attributability on the other, reveal two types of responsibility (sect. 2.2; Talbert 2016, 48–49).

Specifically, Talbert (2011, 151) sees that there are likely to be *degrees* of responsibility and different *forms* of responsibility that are not properly noted. Specifically, for example, if Alessandra deliberately neglected Sheba or adjusted her actions in ways that implied her not caring about Sheba, the degree of responsibility that she would or should face would likely be altered from her merely forgetting about Sheba due to a cognitive glitch (assuming responsibility would play *any* role in the latter case). Further, it seems that the case of Alessandra overall is very different from someone like amerika (the bank robber in the example case *Bad Weather*; see sections 3.3.2 & 6.1.1.3): Alessandra is deprived of awareness of all considerations about her wrongdoing, whereas amerika seems to have considered his actions a lot but has concluded them to be morally justified. Thus, there seems to be nuances of degree and form that are not considered when responsibility is understood as a single phenomenon.

Further, Tognazzini (2010) notes that responsibility may, to some significant degree, be context dependent in the sense that the consequences we are pondering for the target agent may change our judgments. In other words, what would it entail for us to hold Alessandra or Bob or Jack morally

responsible? Does our friendship hang on the line? Us publicly shaming them? Us not liking their status updates on social media? Or us banning them from some platform or service? All these things, and other consequences we might be pondering, may each entail different forms of responsibility, also implicating different processes of evaluation and different epistemic conditions.

As touched on section 5.1.3, Zimmerman (2009, 258–261) echoes the concern of different forms of responsibility, by referring to the possible differences between his “appraisability” and Scanlon and Smith’s “attributability” (or “answerability”). Furthermore, much like Tognazzini and Talbert, Zimmerman (2017, 229–230) argues that *the way* agents are responsible vary a lot depending on their situation, even though most philosophers tend to focus only on the question of *whether* an agent is (in some general way) responsible. In other words, Zimmerman sees there to be *varieties* of praise- and blameworthiness. For example, given that Alessandra would be held responsible, it would be contingent on further situational factors of whether that (form of) responsibility entails, for example, the evaluator resenting, reproaching, berating, shunning, or punishing Alessandra.

I would add that the evaluation of the target agent also depends on the state of mind of the agents *who perform* the evaluation. They too are likely to be affected by situational circumstances in their own life; for example, have they happened to have been frustrated recently by someone who Alessandra reminds them of, did they just get a puppy and thus currently feel greater empathy towards dogs than they otherwise would, or did they not have their breakfast and are thus more prone to get agitated, and so on. They may even be affected by the colors or smell in the room they are performing the evaluation in (Eskine, Kacirik, & Prinz 2011; Inbar, Pizarro, & Bloom 2012a; see also Inbar, Pizarro, Iyer, & Haidt 2012b).<sup>193</sup> Furthermore, they are likely to be affected whether the evaluation is performed amidst the negative affordances of in-group-influenced social media versus one-on-one face-to-face encounter or something in between (sect. 1.1 & 1.2; Crockett 2017a; Schroeder et al. 2017). Further, they are affected by their situationally varying motives (Ditto et al. 2009).

Given these kinds of considerations, it has been suggested that one conception of the epistemic condition could, in some sense, still be the right one (Levy 2005; A. M. Smith 2012). However, the more popular view appears to be that different conceptions capture different aspects of our practices concerning responsibility and there are thus various epistemic conditions applying only to specific types of responsibility. (Rudy-Hiller 2018, ch. 4; see also sect. 5.4.)

## 6.5.2 The pragmatic view as a predominantly normative view

Following this line of stepping back and looking at the arena of discourse from a more distant vantage point, there seems to be three approaches to how the pragmatic view – or any account of the epistemic condition – could roughly be thought about. Namely, it may be thought as (1) one descriptive account of an epistemic condition among others, all more or less equally valid but concerning different types of responsibility or evaluability; or as (2) the one descriptive account of the epistemic condition, superior to all other accounts; or as (3) a normative account on top of any descriptive account(s) we would like to give. By now, it seems quite clear what variety we are dealing with: the very variety we set our eyes on since the pragmatic turn in section 6.1.3.1.<sup>194</sup>

If thought of as (1), the pragmatic view may seem to fit together with some of the accounts examined but not with others, as illustrated in section 6.4.3.3. Alas, this approach of thinking about the pragmatic view would seem misguided, as the majority intuitions do not seem to match with it. Also, the view's primary purpose, in the first place, is to demonstrate its fit with our moral goals (especially compared to Sher's FEC). For the same reasons, it seems even more misguided to think of it as (2). Thus, it seems best to think about the pragmatic view as (3); as a normative account that can be taken to be agnostic about any descriptive accounts, even though it does appear to match the intuitions of some minority (see sect. 6.1).

At the same time, the pragmatic view can *coincidentally* fit together with some descriptive accounts that may be seen to, perhaps unwittingly, follow a normatively praiseworthy account of the epistemic condition, at least in the case of Bob and Jack (see previous sect. 6.4.3.3). Yet, it also fits together with descriptive accounts that are not normatively praiseworthy – e.g., Sher's FEC – in the sense that it remains agnostic of what account(s) best describe people's intuitions. Of course, it can also fit together with other normative or mixed views (Zimmerman's view might be described to at least have significant normative implications; sect. 5.4.3; 6.4.3.3n190), or it can conflict with some.

Furthermore, it is not ruled out that the pragmatic view could not become more descriptively widespread in time, given that our intuitions are malleable to some required degree (see sect. 6.4.3.1). Noting that some minority intuitions would appear to fit together with the pragmatic view – even though partly shaped via pragmatic respect for evidence – it would appear that it is not beyond the capabilities of human intuition, given a certain kind of individual history. It should be noted that the pragmatic view does not deny there might still be degrees and types (etc.) of responsibility, it just points towards what ought to be minimally required for us to evaluate someone as morally responsible if we want to take our moral goals seriously.

## 6.6 Locating the Normative Home Neighborhood

Regardless of whether the pragmatic view could become more widespread, in any case it seems normatively desirable. This is not only because of its individual and societal epistemic and ethical benefits that I have outlined in this overall chapter thus far, but also more broadly because of its potential benefits to the global challenge of morality we have increasingly begun to face. To end the examination of the pragmatic view, and to locate its home neighborhood among normative ethical theories, I will outline this further challenge and how the view may relate to it.

In section 6.6.1, I outline *the tragedy of commonsense morality* that can be viewed as a proxy for the situation we are facing online via polarized moral outrage (sect. 1.1). I briefly consider a utilitarian solution suggested for solving the tragedy, thus considering utilitarianism being a possible normative home for the pragmatic view, and I'll also illustrate how the pragmatic view could alternatively be seen as deontological. In section 6.6.2, I outline a eudaimonistic perspective to the same problem, informed by comparative, fusion, and cosmopolitan moral philosophy (where the first focuses on comparing different traditions; the second focuses on fusing or combining elements from different traditions; and the third focuses on exploring, trying out, listening and speaking, comparing and contrasting, and mostly living at the intersection of different traditions; Flanagan 2011, 1–2). In a sense, this will bring us full circle, back to virtue ethics; locating what appears to be the warmly inviting home of the pragmatic view. Section 6.6.3 concludes the examination with both a brief mention of how these normative considerations could be fruitful to apply to the contexts of artificial intelligence and social media platform design, and with presenting two brief caveats for the overall argument that has been presented in this chapter.

### 6.6.1 Deep pragmatism and the tragedy of commonsense morality

In a somewhat similar vein to the pragmatic view, moral psychologist Joshua Greene (2013), who also has background in philosophy, has argued for what is essentially a rebranded form of utilitarianism that he calls *deep pragmatism*. Greene's context of discussion is in trying to answer what he calls *the tragedy of commonsense morality*: the tragedy of moral conflict and division that emerges when long lost tribes are forced to try to cooperate after having separately developed deeply different moral perspectives (Greene 2013, 1–16). Originally, the tribes can be seen to have developed their distinct perspectives as adaptations to their specific local conditions, for the purpose of solving the prior problem of *the tragedy of the commons*, originally formulated by ecologist Garrett Hardin

in 1968. This prior problem illustrates the problem of cooperation in terms of “Me vs. Us”: the tragedy of resources eroding from a common pasture if self-interest reigns too strongly (19–22). Arguably, commonsense morality solved this problem, to reap the evolutionary benefits of cooperation, even if in cumbersome, unequal, superstitious, and otherwise imperfect ways at times (22–23). However, Greene emphasizes that morality did *not* evolve to promote universal cooperation (26). Thus, as the long-separated tribes are forced to work together with their conflicting moral perspectives, a new challenge emerges: “Us vs. Them” (23–25).<sup>195</sup> By drawing from his dual-process theory of moral judgment, Greene sees that in these kinds of in-group–out-group disputes, our moral disagreements are best resolved by putting our efforts towards setting our (fast and intuitive) ideologies and moral sentiments aside, including deontology, and instead trying to (slowly and rationally) focus on what produces the best outcome for all. This is deep pragmatism: practicing a kind of utilitarian reasoning to find a uniting second-order morality – that is, *metamorality* – that could tame our divisive first-order in-group–out-group passions via seeking common ground not where we think it ought to be, but where it actually is (25–27 & 289–353). At the same time, intuitive thinking may be successfully utilized in the case of “Me vs. Us” -disputes. (see also Greene 2014; for a review of descriptive models of the moral mind, including Greene’s, see Guglielmo 2015.)

However, despite the important problem Greene is addressing, there are doubts – well justified, I think – about whether Greene’s solution of everyone adopting a strictly utilitarian mode of thinking is feasible or even desirable. Often in-group–out-group disputes consist of not necessarily differences of value that could be evaluated in utilitarian terms but, for example, of various relentless self-serving and in-group biases, and variety in interests, beliefs, and interpretations relating to public moral questions. It is not entirely clear if *mere* utilitarianism could resolve all these issues as they would seem to require us to overcome our deeply rooted tribal patterns of thinking in more ways than merely learning to more carefully reflect on our value-differences (cf. Wright 2013, Wright & Greene 2013). Also, not surprisingly, deontologists question Greene’s utilitarianism (cf. Heinzelmann 2018; Nagel 2013). Nevertheless, despite these disputes, it would seem to be a crucial piece of the puzzle raised by the tragedy of commonsense morality that we ought to think things through more carefully and reflectively, while being better metacognitively prepared, whatever the further nuances of the required thinking processes might be or how we might get there.<sup>196</sup>

As the pragmatic view of the epistemic condition (alone) is simply an attempt to arrange public communication in a way to optimize the flow of morally relevant (meta)knowledge and subsequent responsibility – not necessarily to find an all-encompassing metamorality – the comparison to deep pragmatism (or utilitarianism) is not perfect. Still, the tragedy of commonsense morality provides a

good point of comparison for the motivation behind the pragmatic view; a proxy for our predicament (see sect. 1.1 & 1.2). Unlike in Greene's case, however, the result is not necessarily a euphemism for utilitarianism. Alternatively, it could also be understood in a roughly Kantian, Rawlsian, or Scanlonian sense: the aim of the pragmatic view is to strive towards a conversational space we would all like to be a part of when we are initially not aware of how our own epistemic state compares to how things actually are. We should want that those who are less privileged epistemically to be proactively helped rather than counterproductively blamed. Instead of utilitarianism, it could thus be understood that the pragmatic view attempts to describe how we would approach public communication if we were to act in good will and satisfy the categorical imperative.<sup>197</sup>

Yet, given the absoluteness of the categorical imperative (cf. Greene 2013, 331–333), contrasted with the pragmatic view's willingness to doubt and continue to empirically develop itself, I would be prone to *not* think of it in deontological terms. In fact, I am most prone to think of it in terms of *virtue ethics*, also considering that the pragmatic view pays particular attention to our (meta)cognitive *character* and the wider character of our communication. I feel that what I've been engaged in this overall chapter has been the communication of information pertaining to intellectual virtues that ought to henceforth better guide those moral virtues that relate to evaluating agent responsibility (see sect. 2.1n33).

### 6.6.2 Eudaimonia and naturalized Buddhism

Another comparison to the pragmatic view, perhaps more fitting in the overall solution it hints for the tragedy of commonsense morality, could be the moral- and neurophilosopher Owen Flanagan's call for integration of *eudaimonic* considerations into our normative ethical theories: that is, considerations about the causes and conditions of well-being. Flanagan is especially referring to a *naturalized Buddhist* conception of eudaimonia, but also others, that could be integrated with any other ethical theory (for example, we could advocate a eudaimonistic form of consequentialism with deontological constraints). It could thus be understood that the aim of the pragmatic view is not only to point out normative problems in our possible intuitions but also to encourage a more widespread eudaimonia, via encouraging the cultivation of epistemic and moral virtues (e.g., communicative efficacy and compassion). (Flanagan 2011, 17–18 & 95 & 142–143 & 158–159 & 201–202; see also sect. 6.2.1.)

In some sense this project could still be understood as utilitarian (it is, after all, trying to maximize the flow of morally relevant knowledge in society, so as to best facilitate our moral goals

and intergroup cooperation), or deontological (trying to point towards universalizable principles of communication), but I think it is primarily one pertaining to virtues. I think so because cultivating virtues is what sustains our character to optimally follow any other moral goals we may have. And thus, the project is primarily one pertaining to cultivating ourselves towards eudaimonia that sustains virtues and on which the whole project thus rests on. In other words, this calls for cultivating a whole *way of life* and understanding of the world we inhabit, and that inhabits us, that is more sensitive to our predicament than the pragmatic view and Greene’s deep pragmatism are only partly addressing. (cf. Flanagan 2011, 142–143 & 158–159.)

To see what sort of eudaimonia this might more broadly be, Flanagan provides helpful definitions for both Aristotelian and (naturalized) Buddhist eudaimonia. I won’t go into details on the former, as I already outlined some key ideas of Aristotle’s thinking in section 2.1 (incl. 2.1n33), but its definition provides an illustrative comparison to Buddhist eudaimonia (for an explicit comparison, see Flanagan 2011, 155–163 & 165–202). I have adopted Flanagan’s recommended style of using superscript to distinguish between different traditions of approaching the idea of eudaimonia as the good life that everyone seeks but often disagree on what it is, and likewise to distinguish between the resulting feeling states that those different ideas of eudaimonia can produce (Flanagan 2011, 94–95). The focus here is on the latter:

“Eudaimonia<sup>Aristotle</sup> = an active life of reason and virtue where the major virtues are courage, justice, temperance, wisdom, generosity, wit, friendliness, truthfulness, magnificence (lavish philanthropy), and greatness of soul (believing that one is deserving of honor if one really is deserving of honor).” (Flanagan 2011, 95; see also Aristotle & Reeve 2014; sect. 2.1n33.)

“Eudaimonia<sup>Buddha</sup> = a stable sense of serenity and contentment (not the sort of happy-happy-joy-joy-click-your-heels feeling that is widely sought and promoted in the West as the best kind of happiness), where this serene and contented state is caused or constituted by enlightenment (bodhi)/wisdom (prajna) and virtue (sila, karuna) and meditation or mindfulness (samadhi). Wisdom consists of deeply absorbed (intellectually and meditatively) knowledge of impermanence, the causal interconnectedness of everything, that everything (buildings, plants, animals, stars) lacks immutable essences (emptiness), and, what follows from these, that I am anatman, a passing person, a person who passes, a process or unfolding that is known by a proper name, but that changes at every moment, until it passes from the realm of being altogether. The major virtues are these four conventional ones: right resolve (aiming to accomplish what is good without lust, avarice, and ill will), right livelihood (work that does not harm sentient beings, directly or indirectly), right speech (truth telling and not

gossiping), right action (no killing, no sexual misconduct, no intoxicants), as well as these four exceptional virtues: compassion, lovingkindness, sympathetic joy, and equanimity.”<sup>198</sup> (Flanagan 2011, 95; see also Flanagan 2011, 27–29 & 32 & 140–143.)

This definition of eudaimonia<sup>Buddha</sup> describes (1) the kind of mental state that Buddhism offers, that is a stable sense of serenity and contentment, i.e. happiness<sup>Buddha</sup>; and (2) the form of life that promises to produce this mental state, that is living a life of enlightenment/wisdom, virtue, and meditation or mindfulness, being eudaimon<sup>Buddha</sup>.<sup>199</sup> Flanagan views this definition to be an ideal type: the shared core conception of flourishing across all or most forms of Buddhisms prevailing today, including those that reject any dualistic, supernatural and superstitious elements that could be found around this core in some varieties.<sup>200</sup> Below, this is what the Buddhist tradition refers to. (Flanagan 2011, 14–20.)

To further outline the Buddhist exceptional virtues, they are the *Four Divine Abodes* (or the four immeasurables, sublime attitudes; lit. “abodes of Brahma”; *Brahma-vihara*) that are, in part, to be cultivated via meditation or mindfulness (like all aspects of eudaimonia<sup>Buddha</sup>), and realized in action. Compassion (*karuna*) aims to end the unsatisfactoriness or suffering (*dukkha*<sup>201</sup>) of others; loving-kindness (*metta*) aims to bring satisfaction or pleasant feeling (*sukkhā*) or happiness<sup>Buddha</sup> to others in the place of *dukkha*; sympathetic joy (*mudita*) is joy at the success of others or of their good fortune (even in zero-sum games); and equanimity(-in-community) (*upekkhā*) is taking the good of another as its object, i.e. care and concern for all sentient beings, or minimally human beings. The Buddhist tradition sees these, along with the conventional virtues and more broadly eudaimonia<sup>Buddha</sup>, as the currently prevailing fallible lessons that past experience and experiments have discovered to be, thus far, best conducive to genuine feeling state of happiness<sup>Buddha</sup>. Any person who cultivates these exceptional virtues along with other aspects of eudaimonia<sup>Buddha</sup> has entered the *bodhisattva path*: a path leading from our first nature towards a normatively desirable second nature, towards buddhahood. (Flanagan 2011, 12 & 29–31 & 105–109; note 198; see also Flanagan 2017, 239–242 & 310n13.)

Essentially, eudaimonia<sup>Buddha</sup> puts focus on living a compassionate life of contemplation, devoted to others and their well-being at least as much as to your own. Emphasis is put on recognizing oneself as a selfless person, anatman, metaphysically, and thus helping one see that one has reason to be less egoistic and selfish morally.<sup>202</sup> In other words, the focus is on us recognizing our common humanity and connection with all of nature, including each other, and hence seeing first what unites us rather than what divides us. The Buddhist tradition does not, however, usually enforce itself on

anyone, but more so provides a friendly suggestion for exploration if one wants to seek the good life.<sup>203</sup> Conveniently, the exploration need not necessarily start with any text – although it can help – but within oneself (that is ‘no-self’, anatman): to come to notice and understand the nuances of one’s mind through practicing mindfulness or meditation is to acquire greater control of those nuances and thus greater chance of success on one’s path towards eudaimonia<sup>Buddha</sup>. These practices aim to amplify wholesome ways of feeling, thinking, and being, while reducing, ideally eliminating, *the afflictions of the mind*.<sup>204</sup> In the process, they help in approaching the realization of impermanence, interconnectedness, and possibly emptiness, and thus of oneself as anatman (see note 198 above). Nourishing enlightenment, the four conventional virtues, and the four exceptional virtues via mindfulness and meditation essentially amounts to (no-)self-guided metacognitive moral development, and us coming to see the world and our mind more clearly in our moment-to-moment experience.<sup>205</sup> (Flanagan 2011, 109 & 201–202 & 217n7; see also sect. 6.1.1.3n112; Stanovich 2018a.)

The current form of eudaimonia<sup>Buddha</sup> being the best way to achieve any form of feeling state happiness seems to be an empirical claim. However, regardless whether it can be supported by empirical research<sup>206</sup>, the more important point is this: the conventional ways of thinking, acting, and reflecting in the world that currently prevail in wherever we have grown up in – e.g., “East” or “West”, or this or that continent or country etc. – are not unmalleable, definitive manifestation of the way people can or ought to live together, and there might be better ways. The pragmatic view can be seen as an attempt to formulate one aspect of those better ways, and it happens – whether by coincidence or not – to seem well compatible with not only some Aristotelian virtues but also, and I think more fully, with naturalized Buddhist virtues.<sup>207</sup> Thus, it might be a well worth suggestion to try to cultivate it as such.<sup>208, 209</sup> (see also Flanagan 2011, 113–114.)

Alongside his excursions to non-Western traditions, Flanagan has promoted the value of cross-cultural comparisons to spark the creativity of our moral mind, to find new ways of living together – which ought to be a suggestion of rising value in a globally connected world. While practicing comparative moral philosophy, he has paid particular attention to how there is a lot of cross-cultural overlap between what is considered a virtue or vice. However, he has also paid attention to how we seem to be living in especially angry or hateful times (e.g., Pew Research Center 2016; sect. 1.1), all the while the specific value of *anger* particularly divides cultures. Some traditions consider anger very valuable, when practiced the right way (e.g., the Aristotelian West<sup>210</sup>), while others promote and device strategies for the elimination of it altogether (e.g., Tibetan Buddhists and Stoics<sup>211</sup>). When asked about the negative effects of anger, the former tend to describe unpleasant effects on oneself

(e.g., lingering resentment), while the latter tend to describe interpersonal effects (e.g., separation, alienation, and distance between people). Similar differences can also be found in cross-cultural studies: for example, mediated by *cultural regulation* of antecedent situations and habits, Americans tend to experience anger rather frequently and view it as an expression of mature autonomy and independence whereas Japanese tend to experience it rather rarely and view it as something impolite and immature to be avoided for the benefit of conflict avoidance and relational harmony (Boiger, Mesquita, Uchida, & Feldman Barrett 2013; Kirchner et al. 2018; Kitayama, Mesquita, & Karasawa 2006b). Much in line with Flanagan, it would seem to me that there is a lot the moral ecology of the West could learn from the more other-regarding traditions. Some of those lessons would seem to be for the pragmatic benefit of all intergroup cooperation. (Flanagan 2011, 165–207; Flanagan 2017, 133–281; Dalai Lama 2012, 113–136; cf. sect. 1.1 & 2.2 & 6.6.1; for three fascinatingly different example views on anger, see also Bommarito 2017; McRae 2015; Vernezze 2008.)

Of course, this is not to say that some lessons ought not move in the other direction, nor that beneficial lessons could not be gained from elsewhere: by becoming better aware of different traditions and forms of thinking and being-in-the-world, we can all learn a lot from each other and grow together epistemically and morally, on a global scale.<sup>212</sup> Consequently, this might be the direction to take in approaching the solution for the tragedy of commonsense morality: fusion or cosmopolitan virtue ethics, also incorporating the pragmatic view for the epistemic condition, with an aim to cultivate our character accordingly, towards eudaimonia<sup>An\_Informed\_Choice</sup>.

### 6.6.3 Conclusion, application to social media design, and final caveats

To conclude: It is my contention that this line of inquiry, of finding out [A] what are the most pragmatic intuitions possible regarding the epistemic condition (and other nuances relevant to our communication and more largely to naturalistically respectable eudaimonia), and [B] how to best encourage those intuitions to manifest or actualize on a larger scale, should lead the road towards a future of a globally connected humanity. Essentially, this would guard the formation of larger epistemic and deliberative communities where people would not be judged too harshly or quickly, for the benefit of not only optimum conveyance of (morally relevant) knowledge but also for the benefit of being better able to live together on this planet and beyond. The problems of polarization on social media, outlined in sections 1.1 and 1.2, particularly highlight the importance of this endeavor, all the while more people are entering the world of Internet daily (see sect. 1.1n4). This would imply not only paying closer attention to our communication habits (see sect. 1.1–1.2 & 6.4.2–

6.4.3 & 6.6 & 7.3), and consequently to the pragmatic view along with fusion or cosmopolitan virtue ethics (6.6.2), but also paying attention to developing a new kind of pragmatic design ethics for social media platforms and, relatedly, for designing AI; a kind that pays special attention to the pragmatic view and softly encourages people away from vices and toward virtues (see T. Harris 2017).

I think social media and AI ought to be thought as an important space of politics, in a loosely Aristotelian sense. That is, the algorithms and platforms ought to be designed to function as a (part of a) propitious societal context that, in effect, enables and encourages individuals to obtain virtuous habits and eudaimonia<sup>An\_Informed\_Choice</sup> on a global scale.<sup>213</sup> In part, this would imply the electronic environments subtly guiding our biased cognitive systems in a way that we would become (meta)cognitively more efficient in discerning, thinking about, and paying attention to what is salient and relevant towards that goal (or, conversely, efficiently suspending our cognitive patterns when beneficial). Unfortunately, contemporary social media environments seem to be often designed to encourage vices instead. One would assume this has been done unwittingly; i.e., due to ignorance, insofar as the designers have not been knowledgeable of the kind of information presented in this thesis. (see also Honkela 2017; Lewandowsky, Ecker, & Cook 2017.)

Overall, the pragmatic view may be thought as a meta-perspective to (meta)responsibility. That is, it is critically inspecting our metacognition relating to responsibility, while trying to obtain how we might enhance that metacognition to instrumentally better fulfil what are or at least ought to be the aims behind our commonsense assignments of responsibility. Namely, to better fulfil the aim of affecting people's beliefs and actions in relation to what we consider well justified (yet fallible) evidence that demand people's beliefs and actions to be affected, while keeping the needs of an emerging global civilization in mind. Similar to the virtues in Buddhism, this may be understood as an attempt to tame the poisons in our first nature that inhibit cooperation via the tragedy of commonsense morality (sect. 6.6.2n196).

I would like to end this examination with two brief caveats:

Firstly, it should be noted that it might be that for different cases than those of Bob and Jack, different pragmatic criteria could apply, though I think it to be either unlikely or at most consisting of comparatively rare special cases. Insofar as they were found to not apply in some cases, however, then the presented pragmatic account should be adjusted or expanded accordingly, with a distinct pragmatic account given for the further cases.

Secondly, this contemplation has assumed that moral responsibility skepticism is not accepted, which has nevertheless been lurking at the background since section 6.1.2. In fact, some have suggested that moral responsibility altogether should be abandoned, at least in some dominant forms (e.g., G. Strawson 1994/2013; Levy 2011; Waller 2011; see also Caruso 2018). Nevertheless, it seems, at least initially, that some form of skepticism could be compatible with the pragmatic view, and perhaps better than any of the examined descriptive accounts (e.g., Pereboom 2013; Pereboom & Caruso 2018). However, the precise relationship between these would need a separate contemplation, outside the scope of this thesis. To put it shortly and *very* tentatively (subject to change): Skepticism about responsibility may very well be warranted, strictly speaking, but that does not mean we ought to disregard our impressions of what seems or appears for us to be there, insofar as our impressions are otherwise reasonable and compatible with our best empirical evidence. That is, we ought not disregard social concepts that appear to have an empirical, causal effect in the world, the positive aspects of which are not easily replaced. Namely, we ought not disregard responsibility insofar as it appears to be the most effective way of conveying and sustaining what seem to be important moral norms and virtues for a functioning, well cooperating global society. Thus, taking note of the pragmatic view, resorting to assigning responsibility and blame would appear to suggest circumstances for the target agent where all more benign means to affect them would have been first reasonably exhausted, and there are a multitude of more benign means that we seem to casually and systematically neglect.

## 6.7 Summary and Definition of the Pragmatic View

Now that we have considered the general characteristics of the pragmatic view – including considerations about our intuitions in Sher’s example cases and the problem of intuition (sect. 6.1), the importance of the engaged perspective (6.2), the thought experiment of Bob and Jack (6.3 & 6.4) and its implications (6.4.3), the view’s normative nature (6.5), the concern of the tragedy of commonsense morality (6.6.1), and the view’s approximate normative home in virtue ethics (6.6.2 & 6.6.3) – we may finally give an explicit and more concise, approximate definition, in the vein of the consequentialist view (sect. 2.2). A bit of tweaking has been made here to fill in some voids that were left, with explanations following after the definition:

*The Pragmatic View of the Epistemic Condition:*

Whenever an agent performs an action that we consider breaking a moral norm or displaying a vice, we ought to strive to only approach them with civil, compassionate communication that is optimized via our best understanding of the state of empirical evidence concerning how to best fulfill the aim of helping the agent's epistemic state, and consequently aiming to optimize the flow of morally relevant (meta)knowledge in society. Thus, the epistemic condition of moral responsibility is always to consider the fallible current state of relevant empirical evidence.

Currently, the evidence and accompanying arguments would appear to point to the direction that we ought to consider and hold an agent *negatively morally responsible* if, and only if, they satisfy the voluntariness condition, and when they also satisfy any other possible conditions for responsibility that are independent of the epistemic condition, and they either

- (1) are
  - (a) aware of the (meta)knowledge that the act is overall morally negative when they perform it, and
  - (b) precedingly aware of some (meta)knowledge that could have prevented the act with reasonable effort but that they neglected to utilize; or else
- (2) are unaware of the (meta)knowledge in clause 1 when they perform the act but
  - (a) have later gained awareness of the relevant (meta)knowledge, and
  - (b) have understood it, but
  - (c) without displaying genuine other-regarding sorrow and regret for having performed the act.

Otherwise, if these criteria are not met, the agent is to be considered *not* morally responsible. After the agent has broken a moral norm, but before the meeting of the abovementioned clauses can be confirmed one way or the other, we ought to suspend our judgment of their moral responsibility and focus only on communicating relevant (meta)knowledge to them by optimum means we are aware of and reasonably able to perform, while at the same time being aware of our own fallibility. In a sense, holding someone responsible would only be the last effort to convey the moral norm to them, after all other more benign means to convey it along with any relevant (meta)knowledge would have been first reasonably exhausted.

This formulation will likely always be incomplete, however, as our character and actions and communication relating to responsibility are manifested and interpreted in the real, complex, changing world of which we are a part of, with various kinds of individual variation and external circumstances. Thus, examination of the most current empirical evidence, along with appropriate practice, training, and habituation, is always required. This is why the pragmatic view further

emphasizes the deeply embedded tendency for biases and other errors of thought in all humans, and sees examining scientific evidence and cosmopolitan virtue ethics (and *eudaimonics*) as important remedies, and especially so in a community of humans becoming global. Rather than being a strict edict, it is, in effect, a guideline towards an important societal and communicatory virtue that ought to be followed and cultivated ‘in spirit’, along with more general *eudaimonia*.<sup>214</sup>

Clause 1a of the pragmatic view can be connected to the searchlight view (sect. 3.1 & 5.1 & 6.1.1), and thus be called *the searchlight clause*. Clause 1b may be called *the metacognitive clause*, and it can be connected to the emphasis that the most relevant act is not necessarily always the act we are prone to intuitively think as the most relevant act, because the target agent may or may not have lacked crucially important (meta)knowledge during a preceding act (sect. 6.1.1–6.1.2).<sup>215</sup> These two clauses need to be fulfilled together because: [i] if only the former clause is fulfilled, then it can be that the agent did not possess requisite (meta)knowledge needed to prevent the act, yet if this is the case the knowledge should be communicated to them prior considering responsibility; and [ii] if only the latter clause is fulfilled then it can be that the agent is still ignorant of why the preceding (meta)knowledge should have been utilized, i.e. they may be ignorant of why the subsequent act is wrong, and if this is the case it should be communicated to them prior considering responsibility.

For two corresponding examples utilizing the example case of Jack (see sect. 6.4): [i] if Jack is aware of his communication to Bob being counterproductive when communicating to him, but without preceding (meta)knowledge how to improve his communication (even if it could be as simple as “don’t communicate to him” or “do/avoid X to avoid getting triggered into communicating with him”), we should only aim to communicate that knowledge to him prior considering responsibility; and [ii] if Jack is aware of good communication methods but does not know why they should be utilized for the goal of not being counterproductive when communicating to Bob, we should aim to communicate that knowledge to him prior considering responsibility. But if Jack is aware of both (a) why counterproductive communication to Bob is to be avoided, and (b) better communication methods, yet still acts against this better knowledge that he possesses, clause 1 (a+b) is fulfilled and he can be held responsible for his counterproductive communication to Bob because he then overall *knew better*, and can thus appropriately approach a claim of responsibility.<sup>216</sup> Similarly, for example, only if Bob is aware of both (a) why acts exacerbating AGW are to be minimized, and (b) how this might realistically be done (and aware of all the necessary epistemic thinking etc. required to accept AGW in the first place; see sect. 6.3.3), he can be considered responsible if he then continues to act against his gained better knowledge.

Clause 2 can be connected to Aristotle's regret qualifier (2.1.2 & 6.1.1.4), and thus be called *the regret clause*. Similar to clause 1, *unawareness* of both 1a and 1b are required together because [i] if only 1a was required then it can be that after the agent gains and understands that knowledge they are still unaware of how to prevent the act, in which case sorrow and regret may not be expected as they may view the act as unpreventable and hence involuntary; and [ii] if only 1b was required then it may be that after the agent gains and understands that knowledge they are still unaware of why the act itself is wrong, in which case sorrow and regret may not be expected. This corresponds to the example of Jack above, in that only when Jack learns of both (a) and (b) – after having initially been unaware of them – yet does not display other-regarding sorrow and regret, is he to be considered responsible.

Especially noting the importance of the engaged perspective (6.2), and the cases of Bob and Jack (6.3 & 6.4), and how these illustrate the importance of pragmatically communicating to others instead of holding them responsible, these clauses along with the other descriptions above would seem to produce the most pragmatic effects in cases of unknowledgeable agents (or at least they would seem to best avoid the unpragmatic pitfalls in our common habits of assigning responsibility).

Yet, it should be noted that the pragmatic view does not tell what kind or type or degree of responsibility is or should be in question in any particular case (cf. sect. 1.1n3 & 6.5), merely the minimum requirements for any kind of responsibility. Thus, an account of different kinds or types or degrees of responsibility remains an open question after the criteria in the pragmatic view are met. However, whatever kind it is we may be considering in any particular case, the choice should also be made pragmatically whenever possible. That is, the choice should aim to best enable the target agent to still adopt the moral norm or related (meta)knowledge sometime in the future, even though they are currently failing.

It should yet be added that any agent who does not follow this sort of normative guideline (e.g., Jack; sect. 6.4.1), ought to be considered to break what ought to be an important moral norm as they are displaying a vice of a character who hinders important moral aims in society. Consequently, we ought to also communicate with them accordingly, following the pragmatic view.<sup>217</sup>

To end, a brief note on *positive* moral responsibility is in order, as it has largely been neglected in this examination. Like Sher (sect. 4.2), I am prone to think that there is an asymmetry between negative and positive cases, even in a pragmatic sense. Why I am not especially interested in examining positive cases is due to there being an abundance of misguided negative assignments of responsibility

in the world while there is arguably a comparative shortage of any kind of positive assignments. In such a situation, too much of positive reinforcement stirs little worry in terms of it bringing about negative consequences for our moral aims. Yet, people can be encouraged of morally wrong things, of course, and that is something to avoid.

The important pragmatic asymmetry between negative and positive cases appears to be this: If an agent appears to function in a morally admirable way in some situation without awareness (see Sher's positive example cases in sect. 4.2n83), praising them prior communicating relevant (meta)knowledge would appear to have little or no downsides, and could possibly have upsides (e.g., reinforcing the unconscious praiseworthy actions of the agent). However, when an agent appears to function in a morally poor manner in some situation without awareness, blaming them prior communicating relevant (meta)knowledge would appear to be unproductive/redundant at best and counterproductive at worst (see sect. 6.3 & 6.4). Thus, when assuming that the act of the target indeed has the relevant characteristic – negative or positive – pragmatic caution is called much more in the negative case. This asymmetry would appear to boil down to the asymmetry in how people are prone to respond to character-judging stimuli, with – it would seem – negative stimuli where in-group is implicated being especially prone to produce unpragmatic reactions if the agent is unaware of the relevant (meta)knowledge (6.1.3 & 6.4).

This concludes my more theoretical exploration before providing an applied answer to the research question in the following chapter 7.

## 7 HOW SHOULD WE THINK ABOUT THE KNOWLEDGE REQUIREMENT OF MORAL RESPONSIBILITY?

In the introduction, in section 1.3, I put forth the headline question of this chapter as my primary interest, motivated by the context of moral outrage online and related difficulties in science communication. Whereas chapter 6 can be read to provide a more theoretical answer to the research question, in this chapter I conclude my exploration with a more concrete applied answer: I outline how I think we should, normatively, think about the knowledge requirement of moral responsibility, especially in relation to the motivating background of moral outrage online.

As evident from the contents of this thesis, different philosophers have different answers and angles to the question of how we should think about the epistemic condition. I have been able to examine only some of those answers. However, most philosophers try to attain a purely fixed descriptive account or a collection of accounts (depending on whether one views there to be one or many forms of moral responsibility). Conversely, I think the pragmatic view provides a fruitful, much needed perspective. As it is an evidence based normative view, and flexible to accumulation of further evidence, it would seem to not only better enable the spreading of (morally relevant) (meta)knowledge, but also enable better results in terms of its promise of facilitating the aim of usual moral demands. That is, the facilitation of requisite change in others, especially if there are justifiable reasons for those demands. At least this should be the aim of moral demands – otherwise, when we are holding people responsible, we would seem to be doing not much else but flocking our feathers together with some in-group and reinforcing the in-group patterns of thought and behavior. Whether this polarized in-group versus out-group posturing is descriptive of our current majority patterns of moral psychology is beside the normative point.<sup>218</sup>

The predicament of our descriptive situation is particularly well illustrated on social media (see sect. 1.1). It would seem to be a comparative rarity for us to even try to note whether we communicate relevant knowledge when we are expressing our moral sentiments. It is often as if we are automatically assuming that others share our epistemic state, to some significant degree, and thus much of our communication seems to be doing little else than feeding the echo chamber of some in-group. Furthermore, we seem to be often guilty of the ultimate attribution error, i.e., intergroup attributional bias: when negatively evaluating the actions of members in an out-group, we find it hard to put ourselves in their situational shoes; while our own faults and the faults of members in our in-group we tend to associate with our unfortunate situation, whether it concerns our lack of a proper epistemic environment or other situational factors (see sect. 6.2.1n137). It seems that lack of

compassion often leads to unproductive – even counterproductive – expressions of moral outrage, multiplied by the mechanisms of social media that facilitate the contagion and spread of that outrage. Insofar as we partake in these sorts of habits in our lives, we may be described to contribute to the emerging divisions and polarization in society that has arisen in recent years. (sect. 1.1 & 1.2; see also 6.1.3.)

So, how ought we to think about moral responsibility more broadly, but particularly on social media? What would a useful heuristic look like that could conceivably fit together with some people's intuitions and thus be possible for wider adaptation?

To answer these questions, I first outline, in section 7.1, five relevant scenarios under which we may be under when deciding whether someone (or some group) is morally responsible for a perceived transgression. In section 7.2, I suggest a virtuous heuristic – after the pragmatic view (in its current iteration) – that we ought to strive to follow in our moral judgments the best we can. To end, in section 7.3, I will briefly reiterate and emphasize that even though communication of knowledge has an important role in the virtuous heuristic, there are also further communicative measures that should be equally noted.

## 7.1 Five Possible Scenarios and Our Probability of Accurate Agent Evaluation

When we encounter a perceived violation of a moral norm, what seems to follow immediately is an impulse to evaluate agents connected to the violation. Further, what seems to follow is an impulse to hold the agent whom we consider the author of the violation responsible, either explicitly or implicitly. And when we would like to hold someone responsible for some act – i.e., when the impulse first strikes (for example, when reading an outrage-inducing second-hand account of some events on Twitter) – there are at least five possible ways in which the act of the target agent and their (passive) awareness of relevant (meta)knowledge (including relevant (meta)knowledge concerning a preceding act that could have prevented the act in question) may be connected:

- (1) The target agent *was* (or *is*) aware of the knowledge when performing the act
  - (a) and had (or has) no problem performing the act.
  - (b) and struggled (or struggles) to perform the act due to some difficult alternative choices or some other similar factors present in the situation.

- (2) The target agent was *not* aware of the knowledge when performing the act, but has later acquired the knowledge
  - (a) and doesn't feel other-regarding regret (i.e., they feel that they would not have acted differently had they possessed the knowledge beforehand).
  - (b) and has mixed feelings about the act (i.e., they feel that had they possessed the knowledge beforehand, they would have been faced with a situation similar to (1b)).
  - (c) and feels other-regarding regret (i.e., they feel that had they possessed the knowledge beforehand, they would have acted differently).
- (3) The target agent is not aware of the knowledge (and in some circumstances might even be unaware that they have performed or are performing an act).<sup>219</sup>
- (4) There is no connection. The perception of the agent violating a moral norm is mistaken. This may be either due to our understanding of the situation being insufficient, or it being entirely based on false news. (That is, the act was either performed by someone else, and/or performed in a significantly milder or different form than purported, or it was not performed at all.)
- (5) There is a connection, but we ourselves lack relevant knowledge. This can make the act *seem* to genuinely violate a moral norm (e.g., perceiving someone to intentionally spread misinformation) but only due to an unfortunate instance of our own lack of knowledge influencing our moral judgments. (That is, the act was done and, further, was done by the agent and in the manner we have come to understand, but the moral significance of the act is different than what our own level of knowledge or epistemic state leads us to believe.)

Out of these five scenarios, especially scenario (3) would concern following the pragmatic view in the way laid out in the cases of Bob and Jack in sections 6.3 & 6.4. Scenarios (1) and (2), respectively, concern a situation where the agent already had relevant knowledge during the act or time after the knowledge has successfully come to the attention of the agent, roughly matching with Aristotle's original conceptualization in section 2.1.<sup>220</sup> Further, scenarios (4) and (5) expand the list to include additional possibilities that are present in real world situations when an impulse to evaluate someone first rises.

Assuming an agent is perceived to have acted in a morally negative manner, in Aristotle's account the target agent in (2c) would most likely be pardoned, as would possibly – though not necessarily – in (1b) and (2b). Scenario (3) would unravel into (2a)–(2c) after the agent becomes

aware of the knowledge – but they cannot be properly evaluated beforehand. Scenario (4) would not be applicable even for consideration to be pardoned or not: there is no act performed on the part of the target agent. Similarly, scenario (5) contains no ill act on the part of the target agent. Hence, on Aristotle’s account, the only scenarios where moral responsibility would be straightforwardly justified, would be scenarios (1a) and (2a). While understanding the pragmatic view in the way described in chapter 6, I would largely agree with this account.<sup>221</sup>

However, it is crucial to realize that all these scenarios are always present in our global network, when interacting with each other and evaluating each other’s perceived moral violations. As we are far from omniscient, omnipresent, or omnitemporal – even though the Internet may easily give us the false impression of all – we are thus rarely aware of all the facts of the situation we should be evaluating, and thus we are rarely aware of what the situation, in fact, is. It seems that in many cases, when the initial impulse to evaluate an agent first strikes us, we may be under any of the scenarios (1)–(5). Unfortunately, it also seems that for many people the first impulse is not manifested via motivation to perform a careful evaluation process but rather a straightforward jump to conclusions that lead them to believe they are justified in holding the target agent responsible (much like Sher in his example cases and in the case of Bob, although cases on social media tend to be much more ambiguous still). In such cases of lack in temperance, the odds are against us: we have a 2/8 chance to firmly get it right. Assuming all other things being equal, *ceteris paribus*, that would come down to a success rate of 25 %.<sup>222</sup>

Yet, the success rate may be even worse, considering that, for example, (i) false news spreads up to 70 % farther, faster, deeper, and more broadly on Twitter than true news, encouraged by moral outrage and lack of face-to-face interaction, indicating that people can quite easily join in on someone’s initial uncritical impulse (Vosoughi et al. 2018; see also Brady et al. 2017; Crockett 2017a; Pennycook et al. 2018; Suhay et al. 2017); and (ii) people seem generally prone to blame each other for relative non-issues, and often only to uncritically signal their allegiance to their in-group (see sect. 1.1 & 1.2; Tosi & Warmke 2016); and (iii) control or voluntariness condition would bring further uncertainty to the equation (see sect. 4.3 & 6.4.3.3). Furthermore, closely related to (i) and (ii): (iv) initial outrage of individuals on a given topic tends to grow in (in-)groups – producing *severity shift* in judgments – while initial low level of outrage tends to produce lower levels still (*leniency shift*), thus likely feeding outraged sharing behavior but little mitigative sharing behavior by the group members (Sunstein 2018, 5–6; see also Brady et al. 2017; Graham et al. 2012).<sup>223</sup>

Theoretically, (i)–(iv) could be balanced out by people’s intuitive accuracy for skillful evaluations in *most* morally relevant cases, but it is hard to see how there could be such skills, because

(a) false news spreading so eagerly does not exactly warrant confidence (of course, news being false is not a failure of people's moral intuitions *per se* but rather of their epistemic assessments relating to media literacy, but it does not add confidence in either); (b) our judgments of others seem to be very often clouded by intuitive 'in-group must be protected' -triggers along with accompanying biases (e.g., Balliet et al. 2014; Delamater et al. 2015, 452–460 & 481–485; Ditto et al. 2019b; sect. 1.1–1.2 & 6.1.3); and (c) some descriptive theories of the epistemic condition rely on intuitions that do not seem very accurate let alone productive, and hence not warranting confidence either (see ch. 6).

Still, it may be that the majority *do* perform their due diligence, and would want others to perform likewise, when the impulse first strikes. Indeed, *blaming* is a costly signal and people are prone to regulate it especially within in-group (Monroe & Malle 2019), and moral outrage is likely propagated by a mere loud minority (Cohn & Quealy 2019; DataReportal 2019; see also sect. 1.1n5). Hence, it could be that false news spreads more easily because the majority are careful in their judgments of *anything*, and thus patiently doing their due diligence, resulting in *less* news shared on their combined part, and comparatively much *more* shared on the part of an uncritical minority. But that doesn't seem to be the case either, given that Sher's intuitions would be predominant on an individual level, all the while the severity and leniency shifts would affect the intergroup level, also propagating from online to offline (Druckman et al. 2017; Hampton et al. 2017). Our attention spans in consuming media content having become shorter in recent years and decades also speak against us being willing to consider news with patient scrutiny (Lorenz-Spreen, Mønsted, Hövel, & Lehmann 2019). Yet, even if it was the case that the majority are doing their due diligence, on social media what matters more is the visibility of the impetuously impulsive minority and their consequently rapid influence in the public sphere (sect. 1.1 & 1.2). Any regulation via updated moral judgments that may happen can alleviate these impulses in time, but the time difference allows any possible intergroup damage to be done regardless, and *a posteriori* alleviated judgments are not announced nor spread as eagerly as the initial outrage-impulses (e.g., Brady et al. 2017, sect. 1.1; cf. Monroe & Malle 2019). Of course, it may additionally be that some people share false news merely to present what they judge to be *obviously* false, or just trolling for the laughs – neither of which makes the situation much better.

Overall, I am prone to suspect that our success rate of accurately targeting blame to agents on social media is below 25 %, conservatively estimated, or that is what we see there (i.e., for every assignment of blame that is currently being performed online, it has a below 25 % chance of being properly justifiable, let alone explicitly justified). Of course, given the number of variables that might affect the percentage, it ought to be a question of interest for future empirical research.

## 7.2 Mitigating the outrage

Given the low chance of our initial impulse to get it right, and our quick-tempered, in-group facilitated tendencies to express them, how could we make this situation better? There is no straightforward path, of course, but I suggest a virtuous heuristic that we ought to try to cultivate together (as individuals, families, communities, societies, and noted by educators and journalists and social media platform designers, possibly also browser and operation system designers).

Do note that the heuristic has to do with evaluation of the agent, not the act *per se* (the first impulse has already evaluated the perceived act to be counted as a moral transgression, even though our perception of the act may also be mistaken; an error the discovery of which the following heuristic can also facilitate). It can be understood that the heuristic provides a *metacognitive safeguard* to our intuitive, quickly reacting agent evaluations and that it can be cultivated via, for example, mindfulness or meditation, ideally resulting into compassionate right speech (sect. 6.6.2). In time, with enough practice and habituation, this may be integrated into our System 1 processes, instead of requiring more careful System 2 thinking (see sect. 6.1.3.1; Stanovich 2018a).

In section 7.2.1, I outline the virtuous heuristic; and in section 7.2.2, I present some answers to what might be considered complications in the heuristic. If, and when, the heuristic seems complicated when outlined in writing, I have also compiled a helpful flowchart that approximates it (see [Appendix 2](#)).

### 7.2.1 The pragmatic heuristic

Bringing to applied fruition the pragmatic view presented in chapter 6, it would seem to be best to *suspend judgment* of the target agent's moral responsibility when the impulse or suspicion first strikes us, until we have narrowed down whether the agent fulfils the criteria for moral responsibility.<sup>224</sup> The process of evaluation would approximately take the following form (special attention here given to the epistemic condition).

As our first task, we should find out enough about the situation to rule out the possibility of scenario (4); that is, to rule out the possibility of our perception of the agent breaking a moral norm, and thus of them performing a transgressive act, being mistaken. This could be accomplished to a satisfactory degree by, for example, looking into the case via multiple trustworthy sources while paying attention to the principle of charity for the benefit of the agent when interpreting the multiple sources.<sup>225</sup> If scenario (4) is confirmed, then responsibility is inapplicable; and insofar as the

possibility of scenario (4) remains undetermined, likewise the applicability of responsibility remains undetermined, and hence responsibility remains unwarranted. Ideally, (5) – i.e., our own mistaken knowledge misleading our judgments – would also be ruled out, but given that we do not know what we do not know (nor do we *a priori* know what others know), and further noting that we are rarely aware of our own cognitive biases, that would be too tall of an order at this point (cf. sect. 6.1.3 & 6.4.2.2). After sufficiently ruling out (4), we would then want to either rule out (3) – the target agent not being aware of relevant knowledge – or to unravel it into (2a)–(2c), the agent having become aware of the knowledge after the act and having understood it. Given that the situation was not under (3), and it was not unraveled into (2a)–(2c), we would want to find out if the situation was alternatively (1a) or (1b). To do this all, communication would be required (given that this information is usually not found on popular media channels, much less in a reliable format devoid of biased impulsive judgment).

The task would then be to find out if the target agent was aware of the relevant knowledge, either already during the act or having learned about it afterwards. *Offline*, face-to-face, this is a comparatively straightforward task: Communicate with the agent and see whether they are already aware of it, and, if so, inquire whether they were already aware of it during the act. If they were aware of the knowledge already when performing the act, then we can move on to examine scenario (1), and if they have later come to be aware of the knowledge, then we can move on to examine scenario (2). If they are not aware of the knowledge, the scenario thus being (3), we ought to communicate it to them in a manner that we can be sufficiently confident that they have come to be aware of and understand it. This process can take some significant amount of time through many sessions of conversation or dialogue.<sup>226</sup>

Crucially, sometimes the conversation may reveal warrant to suspect we are under scenario (5) – if the target agent or someone else reveals possibly sound knowledge claims that undermine our initial interpretation of the situation. In case of warrant arising for (5), then either the target agent or we ourselves are mistaken in some crucial way (it may also be that we are under the special case scenario of both of us being either mistaken or equally correct in different ways, or a scenario where the bulk of evidence points towards uncertainty, in which case we should also strive to realize this or a more knowledgeable third-party ought to enlighten us both). To confirm whether the scenario is indeed (5), we would need to find out if we can find evidence to falsify or undermine what we ourselves initially deemed as relevant knowledge. This could be done via discussing with the agent, who would also ideally be following a heuristic of wanting to communicate relevant knowledge to you while being aware of their own fallibility. Alternatively, if (and only if) we possess the necessary

critical thinking, information-seeking and science and media literacy skills, along with appropriately intellectually honest curiosity and analytic thinking skills, we could also conduct the attempt of falsification by ourselves from multiple critical sources.<sup>227</sup> Thus, we could either sufficiently confirm (5), and change our mind accordingly, or we could after further examination rule it out, sufficiently confirming that the mistaken party indeed is the target agent. Then we could continue, on firmer ground, to communicate the relevant knowledge to them, or to confirm they were already aware of it. After we had done communicating the knowledge (in case of (3), unravelling into (2)) or had confirmed the agent to already being aware of the knowledge (in case of (1) or (2)), we could move on to examine (1) or (2).

In some cases, it might be that the agent is incapable of absorbing the knowledge, for example due to lack of cognitive ability or due to deeply rooted unhinging epistemic foundations (see sect. 6.3.3), in which case responsibility should be pardoned (this is a special case where (3) cannot be unraveled; due to the agent's epistemic state or cognitive faculties being such that it is beyond their capabilities to absorb the knowledge and thus essentially out of their voluntary control). Furthermore, in some cases it may be that we cannot properly deduce whether the agent is incapable of absorbing the knowledge, in which case responsibility of the agent remains undetermined (this is a special case where the possibility of unravelling (3) would, for example, require unreasonable time or effort for us to properly explore).

So far, I have focused on the offline setting. However, *online* the situation seems to be more complicated: at any one time, a whole mob of people could be trying to communicate to an agent to see whether the agent is aware of some knowledge, after the mob has perceived the agent to have transgressed. As the agent cannot stretch to all directions, it would be imperative for us to first find out if someone has already successfully communicated the knowledge to the agent, and whether that communication would have given warrant to suspect we are under scenario (5).<sup>228</sup> If they have successfully communicated the knowledge, we do not need to do so ourself. If they haven't – and given there is no warrant to believe we are under scenarios (4) or (5) – then we could put the required effort to communicate it ourself, and consequently move on to examine (2).

A further complication online – and sometimes offline as well – is that we may not be able to reach the target agent to communicate with them, and we have come to learn about their transgression only either via second- or third-hand sources or their first-hand writings or videos, for example. If this is the case, then we ought to direct our well-researched comments (i.e., going back to as close to primary sources and their context as possible), including the knowledge the target agent seems to

lack, via the collective (again, after ideally having ruled out that the relevant knowledge had not already successfully reached them and that suspicion about scenarios (4) or (5) had not been warranted). We can then only hope the comments eventually find their way to the agent. And as we deliver the message to the target agent via the collective, we should simultaneously be committed to more direct communication with the agent if they so choose, for otherwise there is no chance of genuine dialogue, only more so monologue. Furthermore, we should ideally write the message(s) aimed at the target agent in their own language or in a *lingua franca*, otherwise we are also more so committed to monologue rather than dialogue (although, often we may legitimately *solely* aim at communicating to people who are likeminded with the target agent and who do not share their language, or to people the target agent's false knowledge has likely affected or can affect, and who should thus be addressed in addition to the original target agent to prevent spread of misinformation). Of course, whenever we deliver a message *via* the collective, we are simultaneously delivering it *to* the collective, thus committing ourselves to dialogue with other members of the collective, not only the target agent (within some reasonable limits concerning our capabilities, time management, and mental health). Thus, overall, to have ruled out (3), we would have either successfully communicated the relevant knowledge to the target agent or had confirmed the knowledge having been successfully communicated to them by someone else or had confirmed that the target agent was already aware of the knowledge during the act.

At this point, we would have sufficiently ruled out first (4), and then (5) and (3). While having investigated the situation and/or having conversed with the target agent, we would also now be aware of whether the situation seems to be (1) or (2). That is, we would be aware whether the target agent was aware of the knowledge already when performing the act or has only acquired it afterwards. Respectively, we could then move on to find out if the situation was, more specifically, (1a) or (1b), if the situation was under (1), or whether it was (2a), (2b), or (2c), if the situation was under (2).

If the situation was under (2) – or has unraveled to be under (2) after initially being (3) – the task would be primarily to rule out (2c). That is, primarily via examining others' already existing conversations with the agent, or secondarily via conversing with the agent if no prior existing communication suffices, and if we had not already found out, we should investigate whether they feel genuine other-regarding sorrow and regret due to the act. If they do, we have no reason to blame them as they seem to already blame themselves (not to mention the many others who have not adopted the pragmatic view will likely already have unnecessarily and uncompassionately, not to mention possibly counterproductively, blamed them). Thus, we should consider that in the case of (2c), the agent is not morally responsible – and we could also try to convey to the agent that they need not be

so hard on themselves either, because they could not have done otherwise as they did not yet possess the requisite (meta)knowledge at the time of the act (see also sect. 7.1n221).

Conversely, if the situation turns out to be (2a), then the agent should be considered morally responsible and appropriately blamed: the blame would function only as the *last resort* behavior guiding tool to signal to the agent that what they did is in fact against a significant moral norm that they ought to internalize in their behavior lest they provoke the condemnation of others. It *may* also be that there is still a genuine disagreement about values or simply opinions at play that no further knowledge can resolve, but nevertheless based on the values of the evaluator the target agent is morally responsible in (2a) (i.e., the target agent appears to display a genuine moral vice of a character or lack of moral understanding, as they have performed what the evaluator perceives to have been a morally wrong act, and performed it specifically without compulsion and without feeling other-regarding regret after later gaining the relevant knowledge).<sup>229</sup>

Furthermore, if the situation turns out to be (2b), then the agent is potentially in some sense responsible and in another sense not responsible, and a final judgment should consider the situational factors more closely. Perhaps a helpful question one ought to ask *in this context* would be: how would one perform in a similar situation, under the difficult parameters that the target agent was forced to perform under, while also considering the agent's epistemic state? I suspect that in many of these kinds of cases moral responsibility should also be pardoned, especially when erring on the side of pragmatic caution. If we think an agent should have acted differently in a particular case of (2b), we ought to provide justification for why we would have acted differently in the case at hand, and thus help the agent and others to guide their future behavior in similar situations (the justification would likely be new (meta)knowledge that the agent lacked during the act). Taking note from Aristotle (see sect. 2.1), we may also describe the case of (2b) in terms of how it would seem prudent to examine how a (more) virtuous character would have ranked the multiple choices in the situation. Thus, the more Aristotelian formulation would be that those who think they would have acted more virtuously in the situation ought to convey the reasons why their way of acting would have been more virtuous, and this can help others to internalize and emulate those reasons in similar situations in the future. Responsibility in the case of (2b) would thus be called for only if it is reasonable to conclude that the agent unknowingly ranked the multiple choices incorrectly, or otherwise unknowingly acted in vice, and yet does not feel other-regarding sorrow and regret after coming to acknowledge that conclusion (see also sect. 7.1n221).

If the situation was under (1), then, in terms of the final resolution, (1a) would correspond with (2a) and (1b) would correspond with (2b). Thus, in the case of (1a) the agent would be responsible,

while in the case of (1b) responsibility would be decided by considering the situational factors more closely.

Finally, it is important to note that even in cases where moral responsibility is determined to be pragmatically justified, some kind of door for possible redemption ought to always keep open. Although it may not often seem like it, people – as the unfolding psychological processes as they are – are capable of deeply changing their views, given enough time. Thus, for example, after learning the relevant knowledge following an unwitting transgression (in the case of (2a)), it can – in some cases – take a long time before appropriate sorrow and regret follows. And even if the transgressor knowingly broke a moral norm with no hinderance whatsoever (in the case of (1a)), the unfolding of future events can bring about significant change in their attitude and character. Thus, even though complete forgiveness for a past act might be out of the question in many cases, willingness for good faith for the future ought to be retained. Essentially, the person may not remain the same as they currently appear, for life experiences in this unfolding universe are an unforeseeable multitude and molder of character.

### 7.2.2 Some answers to possible perceived complications

Even though a critical question could be proposed concerning what qualifies as “relevant” knowledge in any particular case (and what counts as “knowledge”), and thus what qualifies as “relevant knowledge having successfully been communicated”, the definition – including whether it is factual or moral knowledge or both – is secondary in importance compared to people following the heuristic with whatever *they* judge the qualificatory criteria to be as intellectually honestly as they can, while also being cognizant of their own fallibility (and biases where possible). The target agents’ evaluations of whether they lacked some relevant knowledge ought to be taken seriously (along with their possibly differing criteria to ours as the evaluators). What counts as “relevant” (or “knowledge”) in any particular case may also be thought to be dependent on further (meta)knowledge (i.e., philosophical arguments) that ought to be civilly discussed, if needed, while keeping in mind that other people may disagree, at least initially, because they do not share our epistemic state. Insofar as there can be found relevant up-to-date consensual scientific knowledge concerning any particular dispute, it should generally be respected – but for some agents this too may require (pro)active communication and justification of the scientific method and related experts as reliable indicators of arguments that ought to be respected (see sect. 6.3.3).<sup>230</sup>

Another complication that may arise is how to interpret when an agent sincerely feels sorrow and regret for their act after coming to acquire the relevant knowledge. Even though a master manipulator<sup>231</sup> may be very skillful in faking a heartfelt apology, for example, we ought to give everyone who seems sincere the benefit of the doubt (insofar as there is no evidence to contradict them). Basically, we ought to do unto others who seem to have erred as we'd likely wish would be done unto us when we are perceived to have erred. On the off chance an agent is bluffing, we would at least know that if they transgress in a similar manner again, they would then have been aware of the relevant knowledge (given that no new overruling knowledge emerges and that their cognitive faculties, like memory, continues to function properly). Hence, for example, an apology that seems sincere ought to raise forgiveness of one's error (given that they initially lacked relevant knowledge).

Finally, it should be noted that there may be cases where the pragmatic heuristic ought to be skipped – at least temporarily – due to some overruling reasons for which the heuristic is too slow to react. For example, if an agent is holding a gun to someone's head and saying we need to hold them morally responsible for something we know they did not do, lest they pull the trigger, we probably should appear to comply until the situation is under control. But outside of these kinds of special cases, the heuristic would apply.

Admittedly, too complicated a normative heuristic is not very useful as it seems cumbersome to internalize. Therefore, I have sketched an approximate flowchart of the steps laid out above, while additionally roughly accounting for the control or voluntariness condition: see [Appendix 2](#).

If we could cultivate the adaptation of this kind of heuristic in our moral judgments more broadly, and further develop it where need may be, I think we would be doing a service for public communication, not least on the Internet. The key aim of the heuristic – and the pragmatic view in general (ch. 6) – is to cultivate our humility amidst our fallible yet impulsive epistemological and moral judgments concerning the transgressions we perceive, while also providing encouragement for civil, honest and sincere dialogue. Further, it aims to constrain false accusations (7.1) that may only increase divisions in society (1.1), while encouraging for the recognition of the problem of our echo chambers and how different chambers (or individuals within) can genuinely lack important knowledge that we do not, and vice versa (1.1n7).

Crucially, any amount of adaptation helps – this is not an everyone or no one suggestion (although the more the merrier). Ideally, this would result into a heightened sense of equally shared responsibility within the human collective to compassionately help those who are sincerely

unknowledgeable via the most pragmatic communication we are aware of and individually able to reasonably perform (e.g., within the limits that our mental health and endurance permit<sup>232</sup>). While helping us all cultivate epistemic and moral virtues, the heuristic may also be of particular interest to social workers, therapists, teachers, peace negotiators, and similar professionals: not following the heuristic can result into falsely targeted blame, even a circle of blame, and thus the target agent (or group of agents) and/or the evaluator (or group of evaluators) feeling unnecessary distress and lack of compassion. (see also sect. 6.6.)

### 7.3 Other Aspects of Virtuous Communication

As communication and dialogue play a key role in the pragmatic heuristic, I feel it to be paramount to briefly emphasize that the heuristic is only a product of relatively high level of abstraction. That is, it tells the outline, but leaves open the nuances of what virtuous communication might more broadly look like. Indeed, the flowchart in [Appendix 2](#) encourages to constantly develop communication strategies via, for example, communication research and practice. We are not efficient communicators by default.

What might more specific aspects of virtuous communication look like is an interesting question, and I have briefly touched on some suggestions throughout the thesis (see, e.g., sect. 1.1–1.2 & 6.4.2–6.4.3 & 6.4.3.1n197–199 & 6.6; for a helpful psychological theory and path model of blame, see also Malle, Guglielmo, & Monroe 2014, especially pp. 171–174). These aspects concern normative characteristics of communication that we ought to strive for when communicating knowledge to each other, for the instrumental benefit of it optimally penetrating our various belief systems and broader epistemic states, while us remaining sufficiently self-critical, and for public communication to best remain civil and constructive. Like the pragmatic heuristic, also these kinds of further nuances of virtuous communication ought to be cultivated together (as individuals, families, communities, societies, and noted by educators and journalists and social media platform designers etc.).

However, due to the already obscene length of this thesis, examination of these more specific aspects of virtuous communication are left to future examinations. In the meantime, I provisionally encourage us to shun inconsiderate, generalizing group-categories; avoid creating animosity where possible; avoid dogmatic thinking; avoid essentialist thinking; distinguish between people and ideas; distinguish between actions and actors; practice compassion even towards worst perceived enemies;

use social media critically and in moderation; and follow the pragmatic heuristic to our best ability, without forgetting personal mental health (see also sect. 6.6.2). And, perhaps, like also some Stoics and scientific skeptics might suggest, study astronomy to put things into perspective (e.g., Marcus Aurelius 2003, 2.17 & 3.10 & 5.24 & 7.47–51 & 9.32–33 & 12.32–36; Sagan 1994, 8–9; see also Sellars 2002, ch. 3e).

For a future examination of what further characteristics of virtuous communication might look like, I plan on compiling an exploratory addendum on the topic online, to <<https://trainingtheelephant.wordpress.com/thesis-addenda/>>.

## 8 CONCLUSION

In this thesis, I set out to examine the epistemic condition (or the knowledge requirement) for moral responsibility. This was motivated especially by our sad state of discourse, exemplified by contemporary moral outrage and polarization on social media, and related challenges in science communication (as outlined in sect. 1.1 and 1.2, respectively). I wanted to figure out *how we should understand the epistemic condition* that seems to be expressed in some unproductive form in public discourse, if not forgotten altogether (1.3).

To accomplish this, I set out to examine the field starting from the earliest known conceptualization of the two conditions for moral responsibility, freedom/control and knowledge, by Aristotle (2.1); followed by describing the general characteristics of recent discussion about moral responsibility (2.2). Focus was then moved to the contemporary discussion about the epistemic condition, specifically. Particularly, to meticulously outlining George Sher's influential critical account of a position implicit in the works of many philosophers that he calls the searchlight view, which argues that there can be responsibility *only* with awareness (3.1–3.2). Having laid out Sher's largely intuitionist criticism against the searchlight view (3.3–3.4), his own formulation – the full account of responsibility's epistemic condition (FEC) – was also outlined (4). Sher's account expands the searchlight view by arguing that there can be cases where there is responsibility *without* awareness.

I further outlined critical responses to Sher along with introducing other contemporary views about the epistemic condition (5); with the condition seeming to have gathered noteworthy attention only within the last 10–20 years, after an apparently long dormancy since antiquity. Michael J. Zimmerman's revisionist Origination Thesis, a qualified version of the searchlight view, was revealed to be especially central to contemporary discussion, having gathered four main lines of argument that attempt to provide an alternative account against it (5.1 & 5.4.2–5.4.4). Angela Smith's attributionism (or answerability; also associated with Thomas Scanlon) was likewise examined (5.2 & 5.4.4.4 & 6.4.3.3). Outlines of other contemporary alternative accounts were only briefly outlined (5.4), the focus being on the abovementioned. It was noted that the prevailing Eurocentric discussion appears to revolve around formulating competing *descriptive* views about the epistemic condition (5.4 & 6.5).

As Sher's account relies heavily on his intuitions about example cases that to me seem unconvincing, the appeal of his intuitions was undermined and the more general problem in trusting intuitions was outlined (6.1; cf. 3.3.2). Emphasis was put on the importance of the target agents' *metacognitive processes* that should not be dismissed, and a pragmatic turn in our thinking was

encouraged (6.1). Reasonable doubt was also presented concerning another key part of Sher's argument, about the natural home of responsibility being the detached perspective (6.2; cf. 3.3.1). It was argued that if we want to optimally affect a target agent with our communication, evaluation of their epistemic state is paramount, and thus examining also the epistemic states of the evaluators is equally important so as to make sure they are properly paying attention to the target agent's epistemic state (6.2). Considering these shortcomings in Sher's account, a novel empirically informed pragmatic view was formulated and presented as an alternative not only to his account, but potentially to all prevailing intuitive and descriptive accounts of the epistemic condition (6.3–6.7). Particularly, I outlined some epistemic challenges that question any impulses we may have to hold someone responsible without awareness (6.3–6.4). Further, and I think more importantly, I argued that the pragmatic view superiorly satisfies our moral aims or goals that should lay behind our usual moral demands, and has the potential to alleviate our current predicament concerning moral outrage online, and may further clarify challenges in science communication (6.4). Specifically, *the pragmatic view puts focus on us always communicating relevant (meta)knowledge to each other prior responsibility, while pragmatically and compassionately noting the epistemic states of others*. It was noted that unlike the contemporary discussion revolving around competing descriptive views about the epistemic condition (or epistemic conditions), the pragmatic view is *normative*, even though it might fit together with some people's intuitions, possibly minority intuitions (6.5). It was also noted that the contemporary discussion is dominated by merit-based views of moral responsibility whereas the pragmatic view is a consequentialist view (2.2 & 5.4 & 6.1.3.1 & 6.4 & 6.7).

Being primarily a normative view, the place of the pragmatic view among dominant normative theories in ethics was examined (6.6). This was done namely by considering Joshua Greene's tragedy of commonsense morality as a proxy for the problem we are currently witnessing via moral outrage online (6.6.1). Unlike Greene's utilitarianism (i.e., "deep pragmatism") or possible deontological framings in solving the problem, a view emphasising the importance of empirically informed fusion or cosmopolitan virtue ethics was presented (6.6.2). Particularly, naturalized Buddhism was introduced as a system with a lot of promise in aiding us cultivate our metacognition to support virtuous habits, including compassionate utilization of the pragmatic view (6.6.2 & 7). The ethical project of continuing to seek and cultivate virtues was emphasized to be of particular importance as humanity is becoming globally connected and as us humans would thus do well to cultivate any habits that can help us better get along with each other across groups (6.6.3). In summarizing and defining the pragmatic view more concisely (6.7), it was emphasised that in the end it is and will likely always remain but a rough guideline for a normative understanding of the epistemic condition and resultant

agent evaluation and communication pertaining to perceived moral norm violations. This is likely so as the world is evolving, unstill and unfolding, with empirical evidence accumulating and contextually honing. Hence, the pragmatic view appears as something that ought to be followed and cultivated ‘in spirit’, along with more general *eudaimonia* that sustains the required ways of being (6.6–6.7).

Putting the more theoretical exposition of the pragmatic view (6) into good use, it was further applied to how we should concretely approach any perceived violations of moral norms we may witness online or offline (7). This was done by formulating a virtuous heuristic, encouraged to be cultivated, and roughly outlined in the flowchart of [Appendix 2](#). Thus, both a theoretical (6) and an applied (7) answer to the research question of how we should understand the epistemic condition were presented. These were critical normative replies to the currently dominant Eurocentric descriptive examinations of the epistemic condition that emphasize our currently prevailing intuitions and practices among only some subsection of the human population. It may well be that we humans are capable of much better than is commonly understood.

Amidst these multifaceted examinations, I both explicitly and implicitly touched on several areas of potential future research that could be fruitful. These are questions that, for example, social and cognitive psychologists, various philosophers, computer and information scientists, and pedagogists could be interested in. An especially noteworthy one is how might the pragmatic view be implemented in artificial intelligence and social media platform design, in such a way as to encourage cultivation of relevant communicatory virtues instead of vices, and support *eudaimonia* (sect. 6.6.3). Currently, the platforms seem far from helping us cultivate virtues, but the potential is there.

Other questions to be encouraged include the following: What kind of communication in relation to responsibility, even more specifically, would be the best for the pragmatic view to endorse (7.3)? What kind of situational variation might there be? What vision of *eudaimonia* might have the greatest benefits for cooperation (or what might be the best options to make an informed choice on a form of *eudaimonia* worthy of pursuing; 2.1 & 6.6.2–6.6.3)? How accurate are our assignments of blame on social media, more precisely (7.1)? How to best teach the pragmatic view or the virtuous heuristic, or cosmopolitan virtue ethics, to people (6 & 7)?<sup>233</sup> How, besides the pragmatic heuristic, could we best cultivate epistemic and moral humility, instead of groupish and dogmatic moral superiority (7)? What descriptive view of the epistemic condition is actually most descriptive of how people on average think (2.2 & 5.4 & 6)? How might further empirical evidence more precisely fit together with the pragmatic view (6.4.2.4n175)? Besides Sher’s, how would other contemporary

descriptive views about the epistemic condition more precisely compare to the pragmatic view (5.4 & 6.4.3 & 6.5)? It is also interesting how Aristotle's original views might more precisely compare with the contemporary views, especially as many contemporaries seem to have largely forgotten the root of virtue ethics (2.1 & 5.4). Further, it is quite peculiar that the epistemic condition seems to have hibernated for practically 2300 years or so. Is it truly the case that there has not been much discussion about it before the last decade or two, or does the field lack in historical awareness, or do I lack in awareness of the field's historical awareness (2)? And why has the focus been on merit-based views on moral responsibility instead of consequentialist views for the last 50 years or so, as Eshleman states (2016, ch. 1; sect. 2.2)? Have we been overtaken by an intuitive punitive or emotive culture, perhaps by a wind blowing from the West?<sup>234</sup>

I am also especially interested in what might a field of study in social and neuropsychology look like if it were to focus on studying *(meta)metaresponsibility* (6.1.3.1 & 6.6.3); that is, studying our metacognitive processes relating to thinking about responsibility, and particularly in the context of our communication habits. I venture to hypothesize that some interesting discoveries could be made. Further, and relatedly, I am also quite interested in the more precise character of the potential *epistemic bystander effect* (particularly on social media), which seems to be a novel concept altogether (6.4.2.1 & 6.4.2.1n163 & n165). Insofar as there is no directly relevant research that I'm unaware of, social psychologists should be encouraged to further look into these matters. I'm also wondering, might there be some psychological and/or cultural variation among people who are attracted to different views of the epistemic condition? For example, how WEIRD are the general descriptive conceptions of the epistemic condition (Henrich, Heine, & Norenzayan 2010)? Or, might there be something particularly rare about people who are currently attracted to the pragmatic view and cultivating it?

Similarly, it is an interesting question to what degree is adopting pragmatic communication possible in the first place if one's intuitions are not at least to some degree *a priori* in line with the pragmatic view. At the very least, I suspect the view to be something that could be taught starting from primary education (6.4.3 & 6.6 & 7.3). For example, effort should be put towards implementing philosophy (especially ethics – particularly cosmopolitan virtue ethics – epistemology, and philosophy of science), moral psychology, communication studies, compassion interventions, and related activities to curricula on all education levels (6.4.3.1 & 6.4.3.1n182–183 & 6.6.2 & 7.3). Moreover, combining these with exploration of sciences that more generally deal with human thinking and behaviour might be just as important, to enhance our self-critical metacognitive processes by coming to recognize ourselves as deeply flawed beings (6.1.1.1n107 & 6.1.3), at least

insofar as we do not strive to cultivate ourselves towards something better (2.1n33 & 6.6.2n198 & 6.6.2–6.6.3). Currently, these would all appear to be needed for the pragmatic view to become viable beyond the few oddballs inclined to suffer through these kinds of theses and who, in addition, find the pragmatic view appealing enough to cultivate.

Overall, based on the examinations in this thesis, there appears to be tension between our current most common intuitions about moral responsibility and our moral goals (6.3 & 6.4 & 6.6). Thus, to respect our moral goals, it appears that the epistemic condition is best understood pragmatically. That is, it is best understood as an important component of our intuitive habits of agent evaluation and assigning responsibility, and which intuitions can likely be affected in various ways by our upbringing, education, situational and cultural context, practice, and habituation, and that should be affected in ways that our current empirical evidence would suggest to optimally benefit intergroup communication and cooperation and effective distribution of (morally relevant) (meta)knowledge. In practice, this implies we ought to strive to cultivate the pragmatic view (6) and follow the pragmatic heuristic (7) in our agent evaluation to the best of our abilities, thus focusing on conveying knowledge to each other and resorting to responsibility and blame only as a last resort. This would appear to have the potential to attenuate the moral outrage present in our contemporary media landscape, alleviate polarization, enhance public and private communication, foster more constructive media conventions, and facilitate intergroup conveyance of morally relevant knowledge in society. Further, these may harbour intergroup trust and respect. In short, the pragmatic view and heuristic have the potential to bring us closer to eudaimonia, on both personal and societal levels (6.6–6.7 & 7.2–7.3).

To end, the realizations made in this thesis can also be framed as follows: When confined individually, we are ignorant beyond our own help. When confined in a group, we are ignorant beyond our groups help. In this sense, *the limits of our knowledge are the limits in the breadth and quality of our connection*. The novel realization attained in this thesis is that these limits are, in part, determined by our habits of assigning moral responsibility. Insofar as I am not somehow mistaken, we should embrace this and start acting accordingly. Specifically, we should proceed to cultivate habits – like utilization of the pragmatic heuristic for agent evaluation (ch. 7; [Appendix 2](#)) – that better mitigate our individual and societal deficiencies. In the contemporary media landscape of moral outrage and intergroup polarization, we seem to have in many places forgotten the power that we do possess. It may surprise us, but cultivation of carefully thought out virtuous habits and heuristics of thinking may help us activate our capabilities of civil reasoning and dialogue that we *do have*. Or at least we can have, so long as we find the wisdom to work towards them *together*.

# NOTES

## Chapter 1

### Section 1.1

1. Or, depending on what metaethical stance, terminology, or level of examination we choose, we may encounter situations where we feel, for example, a *moral law*, a *moral standard*, a *moral principle*, or simply an important *value* has been broken.

2. In different contexts, an agent may also be held responsible in many other terms found in the philosophical literature. In this context, the most noteworthy seem to be prudential, epistemic, legal, and causal responsibility, not one of which necessarily requires the agent to be morally responsible. Prudential responsibility – also sometimes called *nonmoral* responsibility – has to do with someone having acted either foolishly in relation to some personally significant nonmoral goal (negative case), or conversely having acted prudentially (positive case) (Zimmerman 2009, 248). Epistemic responsibility has to do with being sufficiently careful when hearing and before accepting knowledge claims (positive case) (see Corlett 2008; McHugh 2013), and it is commonly understood that legal responsibility has to do with following the law (vs. not following it), and causal responsibility has to do with having been an effective part in bringing something about via one's actions (vs. not having been). (see also Talbert 2016, 6–14.)

3. Directing blame at someone can take many forms. For example, it can manifest as resentment, indignation, condemnation, accusation, criticism, reproach, or denouncement. Relatedly, it can be expressed via various emotions, like anger, sadness, disgust, contempt, or apathy. Especially in the cases of moral as well as legal responsibility, the target agent is often further considered deserving of punishment – in varying degrees of harshness – due to the perceived responsibility. It is essential, however, to distinguish between the agent and their act: in some cases, an act itself may be morally wrong while the agent may still be pardonable (Aristotle & Reeve III.1; sect. 2.1).

4. Internet access and use are quickly becoming ubiquitous around the globe. From the year 2000 to 2019, estimated percentage of world population using the Internet rose from 5.0 % in March 2000, to 28.7 % in June 2010, all the way to 56.8 % in March 2019 (Internet World Stats 2019a, 2019b). In developed nations, the percentages are much higher: for example, estimated penetration in North America rising from approximately 44.0 % to 89.4 %, and in Europe from 14.9 % to 86.8 %, between 2000–2019 (Internet World Stats 2019b). Most users have migrated to participatory social media, and an increasing number of people report getting their news from there (Internet World Stats 2017; Pew Research Center 2017a, 2018a, 2018b). (see also DataReportal 2019.)

This trend is also reflected in the declining state of the newspaper industry employment rates and monetary returns in the US (Bureau of Labor Statistics 2016; Pew Research Center 2018c). At the same time, statistics show a substantial decline of confidence in traditional press within the last decades, particularly in the US but to a lesser degree also in other English-speaking countries (Hanitzsch, van Dalen, & Steindl 2018). Changes in media trust seem to be linked with changes in political trust, possibly fortified by anti-elitism (ibid.). Globally, people are divided on how capable their local news media is in providing unbiased reporting (Pew Research Center 2018a). To summarize some further differences between the old and new media landscapes, I will be compiling an online addendum to <https://trainingtheelephant.wordpress.com/thesis-addenda/>.

5. Relatedly, it has been argued that social pressures can lead to *preference falsification*, where people's private beliefs become increasingly detached from what is communicated in public. This can further polarization by giving disproportionate space for extreme opinions via driving out moderate ones from the public discourse, leading to a false sensation of what the actual preferences of the majority are. (Kuran 1997; Loury 1994.) Consequently, the remaining loudest preferences, no matter how misguided, rarely get properly addressed as doing so has effectively become a taboo. This seems to be one plausible interpretation of what is happening on some parts of social media, on some topics, especially in networks of perceived agreement, also affecting the offline space (see Hampton et al. 2017; Pew Research Center 2017e; Sunstein 2018; see also Cohn & Quealy 2019).

A closely related phenomenon seems to be the *spiral of silence*: a meta-analysis has supported there to be a significant positive relationship between opinion climate perception and political opinion expression (Matthes, Knoll, & von Sikorski 2018). Also the *majority illusion* seems to be related: in a global network, people must estimate behavior of others from a limited sample of friends and those who are the most visible (i.e., often the loudest) in a network, which may lead to skewed perceptions of the overall prevalence of a behavior (Lerman, Yan, & Wu 2016).

6. For examples and further discussion about some recent mass movements online, I will be compiling an online addendum to <<https://trainingtheelephant.wordpress.com/thesis-addenda/>>.

7. Unlike the agent-driven, self-segregating echo chambers, algorithm-driven *filter bubbles* do not seem to be so prevalent as to warrant a serious concern, at least in terms of online news in general (Bakshy, Messing, & Adamic 2015; Borgesius et al. 2016; Flaxman, Goel, & Rao 2016; Haim, Gibbs, & Lu 2017; Möller, Trilling, Helberger, & van Es 2018). Of course, this does not mean algorithms are necessarily benign: the platforms are, after all, optimizing recommendations in general to maximize time spent on the platform (even at the cost of people's mental health), and malicious third parties are constantly trying to exploit the algorithms for their own ends (for example, to increase polarization via activities on fake accounts).

Philosopher C. Thi Nguyen (2018a; 2018b) has made a vital further distinction between the phenomena of echo chambers and *epistemic bubbles*, where the latter may be understood as a rough synonym for filter bubbles otherwise except that they are not necessarily algorithmically produced. These distinct phenomena can occur both at once or independently, but Nguyen views them to have been often misleadingly lumped into a single, unified phenomenon. He characterizes echo chambers consisting of groups of people who undermine and do not *trust* those outside of the chamber, and epistemic bubbles consisting of groups of people who do not *hear* those outside the bubble. The omission of the bubble to not hear differing viewpoints may be purposeful or inadvertent, while the trust issues in the chamber sees outside viewpoints being actively and systematically discredited. Nguyen compares echo chambers to cults, in the sense that they actively alienate members from any outside sources – for example, via discrediting labeling tactics. Consequently, a bubble is relatively easy to burst: just reach them with information that they've missed. But a chamber is ideologically fortified: it is perniciously and robustly resisting outside influence while feeding an epistemically and often morally dogmatic, black-and-white view of the in-group versus out-group setting (see also Nguyen 2018c).

Nguyen (2018b) argues that the failure to make this distinction is what is largely behind the disagreements in research about the extent of the “echo chamber” phenomenon (cf. Dubois & Blank 2018; Flaxman et al. 2016; Nelson & Webster 2017). Understood via his distinction, it seems clear that echo chambers are a real problem, and even if it may be the case that only a loudly polarized minority belong to them (Druckman et al. 2017), and even if the chambers are regularly exposed to other viewpoints. The point is that the chamber's strategic and resilient discrediting of other views – to the point of being a strengthening act of group identity (Bail et al. 2018) – is distinct from the mere lack of exposure in an epistemic bubble. Furthermore, as epistemic bubbles can be easy to burst, and can thus be very temporal, it may be found in research that they practically do not exist even if it may in fact be the case that in some places, in some topics, they do (temporally) occur. Moreover, it seems especially important to notice the importance of trust: lack of trust has been connected to susceptibility to ideologically motivated conspiracy endorsement (J. M. Miller, Saunders, & Farhart 2015). A time when society (or some segment(s) of it) lacks trust in others or in societal institutions appears to be an especially fertile ground for the formation of echo chambers.

8. As history demonstrates, we were of course well capable of dividing ourselves already before the Internet. However, it seems likely that national newspapers and religions did *locally* unite us more than the present-day fragmented and *glocal* information chambers on the Internet. Still, for a *global* connection to form – if possible – this seems to be a necessary step (see Uz 2015).

9. The “globalists” and “nationalists” in the divide should be understood as ideal types of groups who are trying to hold onto their respective and distinct values (see Haidt 2016; see also Haidt 2012). An agent may also be consistent in their ideas while not distinctly belonging to one category or the other, and there may be distinguished several sub-categories under each (cf. Hawkins, Yudkin, Juan-Torres, & Dixon 2018). In some parts of the world, both sides may find common ground in a benign form of patriotism; a shared sense of identity (Haidt 2016).

10. The average differences between individuals in different political groups may to some significant degree be explained by different underlying psychological factors of individuals, even genetic factors, but being particularly accentuated in the social context we currently find ourselves in (Fatke 2017; Federico & Malka 2018; Hibbing, Smith, & Alford 2014; see also Graham, Haidt, & Nosek 2009; Kanai, Feilden, Firth, & Rees 2011; Verhulst, Eaves, & Hatemi 2012). However, as those differences would be gradual and to some degree variable and contextual – i.e., not strictly black-and-white – and due to likely most people being able to deliberate in propitious contexts and support similar conclusions via different group-specifically appealing premises, this would not mean carefully constructed arguments or other communicative measures could not be effective across groups. Of course, the challenge would be to facilitate those propitious contexts and formulate the group-specific premises to support any arguments we would like to make.

**11.** For example, see Jon Ronson's documentation of several cases of online shaming, and how he connects the phenomenon to the wider history of public shaming (2015a, 2015b, 2015c; see also Blackford 2016).

In terms of fatal reactions to non-existent offences, recently there have been several incidents in India, Mexico, and several other relatively poor countries where lynching mobs have attacked strangers due to rumors spread on social media (Madrigal 2018; Martinez 2018). Furthermore, in the United States, it has been found that the amount and perceived convergence of moral outrage expressed on social media is predictive of politically motivated violence offline during protests (Mooijman, Hoover, Lin, Ji, & Dehghani 2018). Moreover, the violence may be prolonged and also extended to non-provoking outsiders via reinforcement-learning mechanisms associated with the motivational character of intergroup Schadenfreude (Cikara 2015).

**12.** For example, see the case of now resigned Yale professors Erika and Nicholas Christakis (Christakis 2016; FIRE 2016), the case of now resigned Evergreen State College professors Bret Weinstein and Heather E. Heying (Heying & Weinstein 2017; Perrino 2017), the case of now fired Google employee James Damore (Lewis 2017; Singer 2017; see also Stevens & Haidt 2017), and the case of Muslim reformist Maajid Nawaz (Nawaz 2016b; Southern Poverty Law Center 2018; see also Chandler 2019; Nawaz 2015, 2016a; Reeves 2019).

**13.** As the popular term "fake news" has been politicized to function as a catch-all-term to refer to *any* news or news source someone subjectively holds in suspect, Vosoughi et al. (2018) encourage the use of the terms "false news" and "true news" that more straightforwardly refer to the veracity of the information presented and do not implicitly assume anything about the intent. Also, they use a broad definition of "news"; referring to any story or claim with an assertion in it.

In the study, the veracity of ~126,000 stories were examined along with the tweets that shared them. Their veracity was evaluated via information from six independent fact-checking organizations, and the stories were tweeted by ~3 million users more than 4.5 million times between 2006–2017. When controlling for the account age, activity level, the number of followers and followees of the original tweeter, the verified status of the account, and even for bots (via utilizing two state-of-the-art bot-detection algorithms prior analysis), retweeting falsehoods was found to be 70 % more likely than the truth. As the researchers utilized numerous diagnostic statistics and manipulation checks, the results seem robust. Additionally, novelty value of the information shared was analyzed and found to be a highly likely factor in motivating the sharing behavior.

One can easily see how this may all result into online information cascades containing a significant number of false attributions of moral responsibility. As even a single outrage-inducing headline can spread overwhelmingly quickly, far and wide, thus even a singular story – regardless of veracity – can have a vastly disproportionate effect to our view of the world.

**14.** Caution and further research is called for, though, as the attenuating effect of analytic thinking may only be confined to very specific circumstances – for example, to a setting where utilization of analytic thinking in relation to 'news headlines' is 'prompted' (cf. Pennycook & Rand 2019b, sect. 5.1) – and as at the same time it does not appear to attenuate but actually *exacerbate* ideologically motivated reasoning and thus political polarization in other contexts such as climate change, gun control, and selective exposure to political information (cf. Kahan 2013; Kahan et al. 2012; Kahan et al. 2017b; Kahan, Jenkins-Smith, & Braman 2011; Knobloch-Westerwick, Mothes, & Polavin 2017; see also Ditto et al. 2019b; Kahan 2017b; Van Bavel & Pereira 2018).

Still, a more analytic thinking style has been further linked to, for example, skepticism about epistemically suspect beliefs like religious and paranormal beliefs, skepticism about conspiracy theories and alternative medical claims, making less emotional or disgust-based moral judgments, being less cooperative and more rationally self-interested in social dilemmas, and being less receptive of meaningless statements designed to sound profound (Pennycook, Cheyne, Barr, Koehler, & Fugelsang 2015a; Pennycook, Fugelsang, & Koehler 2015b). Ståhl & van Prooijen (2018) have indicated, however, that at least some of these sorts of results may not reflect analytic thinking so much as general cognitive ability combined with valuing epistemic rationality. Though the research does not yet seem to exist, this may be the case for the link between analytic thinking and reduced false news susceptibility as well. Further, though not yet researched, it may be hypothesized that general cognitive ability combined with valuing epistemic rationality may be connected with scientific curiosity. This could be because both have been connected to reducing political bias, and because epistemic rationality as motivation to pursue and hold accurate beliefs could be seen to derive from curiosity, at least to some degree (see Kahan 2018; Kahan et al. 2017a; Ståhl & van Prooijen 2018).

**15.** The propensity of humans, as the gregarious social animals that we are, to divide themselves into in-group favoring groups via the most arbitrary distinctions has been well demonstrated via research utilizing the *minimal group paradigm*.

For example, a mere coin flip can form discriminatory group allegiances, both in adults and already in preschool children (e.g., Amichai-Hamburger 2005; Delamater, Myers, & Collett 2015, 456; Yang & Dunham 2019; see also Bloom 2013).

**16.** To bring some balance to these concerns, what might it look like when we occasionally do manage to actualize the *positive* affordances of social media? See, for example, anthropologist Michael Wesch's historical, albeit a bit dated, *Anthropological Introduction to YouTube* (2008). Or, see the introductions to the web's many educational platforms (e.g., Anderson 2010; Khan 2011; Koller 2012; Agarwal 2013), though compare them with physics educator Derek Muller's constructive points (cf. Veritasium 2011, 2014; see also 2veritasium 2017). Wikipedia is another example, albeit less of a social media than an encyclopedia that combines crowdsourcing with hierarchical and democratic structures (see Wales 2005; Wikipedia, n.d.).

In a recent survey in the UK (Royal Society for Public Health 2017), it was found that between Facebook, Instagram, Snapchat, Twitter, and YouTube, *only* YouTube had a net positive impact on the self-reported health and well-being of the 14–24-year-olds surveyed (the most active age group to use social media). Insofar as these results can be further substantiated, this could be hypothesized to have something to do with the distinctive nature of the comparatively long-form video format, where communication is not reduced to mere text, photos, nor short snippets of life highlights (Schroeder, Kardas, & Epley 2017). However, caution should still be encouraged, especially in the case of young children. For example, it has been noted that autoplaying cartoons or nursery rhymes on YouTube may bring up material not suitable for young minds (Maheshwari 2017; see also Pew Research Center 2018e). Content on YouTube, like on many other platforms, consists of a whole spectrum of human endeavors, and many of them unique to the hard-to-control Internet culture(s).

To be clear: There are many *incredible* educational, therapeutic, deliberative, creative, crowdsourced, and otherwise pro-social channels and communities on the Internet, and within various social media platforms. Similarly, the related technologies, algorithmic environments, and AI's information of ourselves can also be harvested for many valuable purposes. The worry presented in this chapter concerns the current level of imbalance between the positive and negative affordances that are currently, on average, actualized on the most popular social media platforms. And it concerns our relevant behavior relating to moral responsibility, also echoing within the larger society, at least in North America and Europe.

## Section 1.2

**17.** For an overview of the many positive trends in human progress, and how we are nevertheless – at least currently – highly biased towards the negative, see *Factfulness* by statisticians Hans Rosling, Ola Rosling and Anna Rosling Rönnlund (2018). Another optimistic overlook has been recently outlined by psychologist Steven Pinker in his book *Enlightenment Now* (2018); and further from a different angle by physician and sociologist Nicholas A. Christakis in *Blueprint* (2019).

**18.** Of course, insofar as those who do not believe in AGW are in a closed echo chamber no matter the distribution of political affiliation among group members, that is a problem in itself.

**19.** In Europe, climate change is not as politically divisive a topic (European Commission 2017). In 2015, in 36 examined nations (not including China), party identification was the strongest predictor of climate change concern in Australia, the USA, Canada, Germany, Israel, and the UK, while in 27 countries partisan differences were not statistically significant predictors. Commitment to 'democratic principles' was found to be the most important predictor everywhere except English-speaking Western democracies. (Lewis, Palm, & Beng 2018.) Another 25-nation study has suggested that political culture in contemporary US, in particular, may be a comparative anomaly in terms of the strength with which it inspires conspiratorial ideation or conservative ideologies to be directed towards questioning AGW (Hornsey, Harris, & Fielding 2018a). Unfortunately, that culture yields a lot of global power.

Furthermore, in 2015, in 40 measured nations, concern about climate change was best noted in Latin America, Africa, and Europe – with the USA and China on the opposite end. The global median for those who considered it a *very* serious problem was 54 %, and the median for those who supported limiting greenhouse gas emissions was 78 %. (Pew Research Center 2015.)

Overall, there is still a lot of room for improvement both in terms of public awareness and political decision-making. According to a meta-analysis published in 2015: Over the past quarter century, concern about the issue among developed nations has practically stayed the same. In the US, not counting occasional fluctuation, people worrying about the issue 'great deal' or 'fair amount' has stayed around 60 % since the end of 1980s. In Europe, there has been little aggregate change since early 1990s. In parts of Africa, developing Asia, and Latin America, perception of the seriousness of the

issue has notably increased in late 2000s. As a global aggregate, considering climate change a threat has remained practically unchanged since 2000s (earlier global data is scarce). However, it could be argued that measuring “concern” or “worry”, for example, may be misleading: it is not necessarily an indicator of how noteworthy one takes the issue to be, as it may also function as an indicator of attitude amidst recognized adversity. Still, after the meta-analysis in 2015, there has been some cautiously promising developments in the US: the amount of people who worry “great deal” has risen from 32 % in 2015 to 43 % in 2018. (Capstick, Whitmarsh, Poortinga, Pidgeon, & Upham 2015; Brennan & Saad 2018; Saad & Jones 2016.)

**20.** Note that AGW is only one scientific question stirring within the public sphere. When examining popular questions more broadly, it becomes clear that neither of the major parties in the US is the party of science. They both possess their own areas of scientific inaptitude, seemingly guided by what best fits with their political beliefs or agenda at a given time (i.e., what fits with their respective sacred values). Cognitive scientist Keith E. Stanovich – who in recent years has been developing the rationality quotient (RQ) – has described how both sides are just as ‘rational’: for example, while more Republicans tend to suspect AGW and evolutionary theory, more Democrats apparently seem uninformed about, for instance, the lack of bias in hiring women for tenure-track university positions in STEM disciplines and of biological evidence about sex differences (Stanovich 2017; see also Nisbet, Cooper, & Garrett 2015; Washburn & Skitka 2017).

Further, a meta-analysis has indicated that partisan bias among liberals and conservatives is overall symmetrical (Ditto et al. 2019b; though for a critique and an answer, see Baron & Jost 2019; Ditto et al. 2019a). However, another meta-analysis has indicated that there is motivational cognitive asymmetry between the two: for example, on average, conservatives in the US tend to be more cognitively rigid and dogmatic, and less tolerate of ambiguity (Jost 2017; though for a partial critique, see Ditto et al. 2019a; see also sect. 1.1n10). In any case, despite these ongoing disputes, both political sides seem to have important things to say that the other side is often oblivious to (Haidt 2012, 319–366).

Unfortunately, it seems the obliviousness about the other side may to some degree produce notable ideological bias even to research in social sciences (Duarte et al. 2015; Eitan et al. 2018; Haidt 2011a). This can happen due to lack of viewpoint diversity and poor bias-reduction mechanisms – noting that an overwhelming majority of social and psychological scientists, at least in the US, identify as liberal, whereas the populations they study are of course ideologically much more heterogeneous (ibid.; Inbar & Lammers 2012). Discussion among social psychologists is ongoing, but at least tentatively it would seem that if the discrepancy of ideological distribution among researchers has an effect on research replicability, it is easily overstated, as it would seem that at most those who hold extreme ideologies are affected (whether liberal or conservative; Reiner et al. 2019). Still, it seems that there is more willingness to discriminate against conservatives in US academia (Eitan et al. 2018; Inbar & Lammers 2012), and political homogeneity may constrain research scope and direction (Duarte et al. 2015; Eitan et al. 2018). (see also Ditto et al. 2019a, 2019b; Kahan 2013; Kahan et al. 2017b; Stanovich & Toplak 2019.)

Briefly looking into the issue of (lack of) viewpoint diversity, one particularly noteworthy observation can be made: Firstly, an increased distrust of science among conservatives in the US (relative to liberals and moderates) seems to have arisen since the early 1990s, while liberal trust has remained steady, even though trust among liberals and conservatives was about equal in the 1970s and 1980s (Gauchat 2012). Secondly, the “liberalization” of academia seems to have begun from the early 1990s (Abrams 2016; Duarte et al. 2015). What is noteworthy is that these two trends *appear* to be correlated (see Abrams 2016, Fig. 1; Gauchat 2012, 175, Fig. 1; Duarte et al. 2015, 3, Fig. 1). Further, in just the last few years, there has been an astonishing negative spike in Republicans’ perceptions of the value of colleges and universities, due to perceived political agenda (Newport & Busteed 2017; Pew Research Center 2017g). To some degree, similar perception about social scientists in particular may even be bipartisan (Hannikainen 2018; Washburn & Skitka 2017).

Overall, also noting the wider “replication crisis” (e.g., Reiner et al. 2019; Shrout & Rodgers 2018): to minimize the chance of us being misled about how our social species works, as well as to not increase public distrust of science in the US or elsewhere, and further to maximize the potential of science communication, it seems that special care is currently called for when designing studies and interpreting results, and especially if these processes could be seen to favor liberals. Further, we should all treat each other fairly and with respect.

**21.** It seems important to note that religiosity and political conservatism, perhaps unlike is usually assumed, play distinct parts in science acceptance, even though they are reliably intercorrelated. That is, for example, in the case of vaccinations, opposition can indeed be best correlated with religiosity and either not correlated or at best only weakly correlated with conservatism (the weak correlation was found only in a smaller pilot study); whereas at the same time, in the case of denying AGW, both religiosity and political conservatism can be implicated, with the latter being the better predictor. Hence, science acceptance is topic-specific, each variance driven by different combination or emphasis between four major predictors that tend to intercorrelate and can thus be easily confounded: *political ideology, religiosity, morality, and knowledge about science*. (Rutjens et al. 2018; see also Graham et al. 2009; Malka, Lelkes, Srivastava, Cohen, & Miller 2012; McKay & Whitehouse 2015.)

22. The items have been, for example, scientific true/false claims or questions, such as “the center of the Earth is very hot” [true], “all radioactivity is man-made” [false], “does the Earth go around the Sun, or does the Sun go around the Earth?” [the former], and “antibiotics kill viruses as well as bacteria” [false] (see, e.g., Kahan et al. 2012).

23. Zhou (2016) provides a potentially conflicting finding regarding the effectiveness of framing AGW in Republican values. He concludes that polarization or politicization of an issue may be a constraining factor in framing effects, potentially even resulting into a *backfire effect*, increasing Republican opposition. However, the way Zhou framed the issue in his study were essentially blunt, short statements of call for action, and further seeming to sell the problems to the participants with solutions that might still go against their deeper values, and without selling the knowledge of causal mechanisms behind climate change or suggested policies at all (cf. Campbell & Kay 2014; Fernbach, Rogers, Fox, & Sloman 2013). For example, Feinberg & Willer (2013, 2015) use more complex framing – for example, descriptive messages of several paragraphs – and produce successful results, albeit they do not explicitly study the topic of climate change. These and further nuances of framing can matter in subtle ways, as Zhou also acknowledges. Persuasion can also take time not available in a study setting, and the right communicator whom the target group can get behind (Wolsko et al. 2016). Still, there may be more resistance to moral reframing at the ideological extremes, which can prolong the process in polarized times – though likely much less than if reframing was not utilized.

24. The success on an individual level, however, is likely contingent on their level of trust in the scientific community. By the same token, trust in the scientific community may reduce polarization on scientific topics, and it would thus also be something to pay attention to and promote. For example, in addition to striving to enhance the conservative image of educational institutions (see note 19 above), epistemology and philosophy of science ought to be promoted, especially when in many places they are currently nowhere to be seen on national primary education curricula.

25. Often knowledge and reasoning abilities associated with science literacy are viewed to be key intellectual capacities that a modern democratic citizen ought to have (J. D. Miller 1998). However, American professor of Law and professor of Psychology Dan Kahan, who is the lead researcher of the study on scientific curiosity (Kahan et al. 2017a), has made a case that science literacy without science curiosity can impede public recognition of the best available evidence and even *deepen* polarization (see also Drummond & Fischhoff 2017; Lewandowsky & Oberauer 2016). This, he describes, is because smart, scientifically literate people who are not scientifically curious will primarily digest arguments by other smart people who are similarly biased as they are, and strongly follow that in-group. But the more a person is scientifically curious (i.e., the higher they score on Science Curiosity Scale, SCS, that Kahan and his colleagues have developed), the more likely they are to go beyond from hearing only their biased in-group to actually pursuing what the best arguments and scientific evidence are. In other words, the more curious someone is, the more motivated they seem to be to pursue information beyond their in-group. Of course, in that pursuit other qualities are also important: for example, science and media literacy, and related reasoning and information-seeking skills. These may all be viewed as important individual capacities and skills that a modern democratic citizen ought to have, and that can be taught, learned and enhanced. (Kahan 2018; Kahan et al. 2017a; see also Kahan 2015, 2016; Kahan et al. 2012; Ståhl & van Prooijen 2018; Taber & Lodge 2006.)

26. This means violence ought to be avoided as well, lest it may backfire (Simpson, Willer, & Feinberg 2018).

## Section 1.3

27. This description of the freedom requirement seems to fit the examinations I conduct. However, there are subtle differences in how different philosophers see and conceptualize freedom, control, voluntariness, and knowledge. The literature overall contains various kinds of relations between the terms, due to various conceptualizations and the related views about how those relate to responsibility. For example, even though the orthodoxy is to see the freedom and epistemic conditions as distinct, some argue that the latter is a component of the former, and some think there is no epistemic condition to begin with (see Rudy-Hiller 2018, n1).

28. A related question may further illustrate my interest in the aforementioned: Should we or should we not consider agent S responsible for something X if the agent does not have awareness of knowledge Z that is crucially relevant to X? For example, should an agent be held morally responsible for global warming if the agent is not sufficiently aware of the relevant scientific evidence and how it relates to their behavior and actions? Or, as a further example: should an agent S living in an ideological silo be held morally responsible for something X if the crucially relevant knowledge Z has not

penetrated that silo? These are the kinds of questions behind my interest in how we should think about the knowledge requirement.

29. For a summary of the general discussion about the different form of knowledge and their relationships, see Fantl 2017.

30. Philosophers generally tend to focus on the negative side. Talbert (2016, 2–3) describes two reasons for this: (1) most philosophers assume that the kind of relation one must bear to bad action in order to be blameworthy is symmetrical to the relation one must bear to good action in order to be praiseworthy; and (2) there is more at stake in the negative cases, as being blamed for something can have *unpleasant* consequences in minimum for the target agent, but praising someone can usually at worst bring about *pleasant* feelings for the target agent. Thus, undeserved blame is a more significant concern than undeserved praise.

31. I generally use and understand “science” in the broad sense: comprised of both the natural and psychological/social sciences whose aim it is to make discoveries and compile a body of knowledge about how reality works and what it is comprised of, including the reality of human mind and behavior as well as social phenomena.

## Section 1.4

32. As one clear exception: my motivation (in sections 1.1 and 1.2) largely emerged after or partly simultaneous to my delving into the topic. Also, relating to chapter 6, I formulated my own view – that is, the general idea of the pragmatic view – before being aware of the conversation by Sher and others, nor being aware of Aristotle’s original formulation.

## Chapter 2

### Section 2.1

33. To provide some background for the following discussion, below is a condensed contextual outline of some key ideas in the first two books. This is interpreted from the dense text of the original source translation; with help of various guides, primarily Pakaluk 2005. For a much more nuanced outline along with some critical commentary and alternative interpretations, see Pakaluk 2005.

In Book I of the *Nicomachean Ethics*, Aristotle describes that all people seek *eudaimonia* (i.e., human flourishing, well-being, fulfilment, or happiness; or the good life; constituting the highest good or goal), but that some pursue it with misguided means. He especially mentions three common but misguided means: [1] life of money-making, pursuing wealth (misguided because it is merely a means to an end, not an end in itself); [2] life of politics or leadership, pursuing honor (misguided because it is merely dependent on others’ opinions, and because honor is secondary to the virtue which receives it); and [3] life of comfort or ease, pursuing pleasure (misguided because pleasure is always fleeting, and one ought not become enslaved by it). (Aristotle & Reeve 2014, I; Kraut 2018, 2; Pakaluk 2005, 47–86.)

Aristotle reasons that to pursue *eudaimonia*, it is best to first obtain the *function* of a human being – i.e., that for the sake of which we humans uniquely exist – so that we could determine our *virtues* (i.e., traits that contribute to someone being a ‘good’ human, or a functioning human; and in this sense, a *beautiful* human) (ibid.; see also Flanagan 2011, 147–148; Pakaluk 2005, 108). Analogously, we may think of how in pursuing the highest good for a house, it helps to realize that the function of a house is to provide shelter, which can enable us to determine what contributes to something being a good house (e.g., solid foundation, strong building material, thermal isolation, etc.).

Drawing from his psychological and biological works, Aristotle reasons that as only humans are, in his view, capable of higher-order rational activity – or actions in accordance with reason (or in accord with reason) – that activity is our unique function [even if this was not unique for humans, it would arguably still be our highest function or at least an extremely important one]. And as virtues are traits that contribute to someone being a good, functioning human, they include and are further determined via actions in accordance with reason. To pursue *eudaimonia* – i.e., to pursue the *highest* human good or being a *well*-functioning human – is thus done via cultivating virtues in accordance with reason. For Aristotle, this is indeed a *practice* that last a lifetime (after one has grown up and learned to pursue *eudaimonia*); much like a well-functioning house serves its function for its lifetime (after having been properly built). Thus, life lived in practicing rational activity in accordance with virtue is what *eudaimonia* consists in. At the same time, things like good friends, family, wealth, honor, and pleasure also remain valuable insofar as they support *eudaimonia*. To some degree

these require *luck*, and hence insofar as it may not be pursued without them, eudaimonia favors good fortune. (Aristotle & Reeve 2014, I; Kraut 2018, 2; Pakaluk 2005, 47–86.)

In Book II (and partly in I.13), Aristotle outlines how he views there to be two kinds of virtues that depend on each other: thinking-related and character-related (i.e., intellectual/epistemic and moral/ethical). He sees them to be acquired in propitious societal contexts. Thus, if one is born into the wrong kind of society or environment, these may be impossible to acquire. That is, our inborn capacity or potential for virtue cannot then be actualized. Hence, Aristotle also emphasizes the immense importance of politics, education, and upbringing in creating a propitious societal context where children can grow up to flourish. (Aristotle & Reeve 2014, I.13 & II; Kraut 2018, ch. 4–5; Pakaluk 2005, 87–117; see also Aristotle & Reeve 2014, X.9.)

The thinking-related virtues are virtues of our *rational* part – our part consisting of practical and theoretical reason (see Aristotle & Reeve 2014, VI; Pakaluk 2005, 219–220 & 278 & 278n11) – and they mostly come about and grow as a result of *teaching*, thus requiring experience and time to learn (like the virtue of rational activity that Aristotle himself teaches to be the unique function of a human being). In contrast, character-related virtues are virtues of our *nonrational* part, and they result from *practice* and *habit*. Good habits ought to be cultivated, while bad habits ought to be discouraged. (ibid.)

Aristotle seems to view that the first step to acquiring a character-related virtue is typically by performing actions like those people who have the virtue, because one is directed to do so by some discipline or rule (e.g., in their upbringing when young, or in legislature when mature, encouraged via honors and punishments), and/or because one is naturally inclined to want to get along with other good people and emulate them. Eventually, one will be cultivated to *want* to perform those actions as they have become a habit that is in accord with practical reason, and one gets pleasure from performing them. Essentially, Aristotle views character-related virtues like any skill: it takes (guided) exploration and disciplined practice, and eventual (voluntary) habit, to cultivate a morally virtuous character, much like it would to cultivate the character of a skillful archer, for example. However, one ought to be careful not to cultivate *vices* instead (i.e., traits that contribute to someone being a ‘bad’, poorly functioning human). (ibid.)

To outline the relationship between character-related virtue and vice, Aristotle describes virtue to be something that resides within a *mean* (popularly known as the *golden mean*), between two extremes of deficiency and excess. For example, courage is the mean between the vices of cowardice (deficiency) and rashness (excess). The mean is not an absolute, however, as it depends on the specific virtue of whether it resides closer to one extreme or the other. For example, courage is closer to rashness than to cowardice, but it is not rashness. Different situational particulars and personal characteristics also affect the mean (for example, regarding the virtue of temperance in terms of drinking alcohol, an alcoholic would generally find their virtue at the extreme of sobriety, whereas others may have more room to find theirs). (ibid.)

Unlike in moral virtues/vices, there is no excess in intellectual virtues as they appear to concern truths and falsities or correctness’s and incorrectness’s or (appropriate) affirmations and negations, and our ideally only increasing knowledge or awareness of them. Intellectual virtues thus also provide the practical standard of action that a virtuous character will follow (for example, my knowledge about alcohol and how my body reacts to it ought to inform how I approach the substance in terms of temperance). One important intellectual virtue is *practical wisdom* (or *phronesis*), associated with the ability to *deliberate* correctly about what is good for oneself and what is good for humans in general. Someone who has learned and cultivated this virtue may be called a prudent or a mindful human. Aristotle views that the prudent human, who has also acquired a moral virtue through practice and habituation, will know in a relevant situation where the mean resides and thus *choose the right action*. And in so doing, they are fulfilling their highest function, and thus living a life of *eudaimon* (that is, living a life of someone who is contributing towards a life of eudaimonia). (ibid.; Aristotle & Reeve 2014, VI; see also Flanagan 2011, 12; Pakaluk 2005, 206–232.)

Practical wisdom seems to also enable what Aristotle considers the most highly ranked virtue (or what may be interpreted from his obscurity as being such); one that is not only good but best for humans as it appears to be what other virtues aid or target at, and is in this sense most complete and goal-like: *theoretical wisdom* (or *sophia*), associated with understanding, ‘scientific knowledge’, and related active contemplation and utilization of one’s mind. (Aristotle & Reeve 2014, I.7.1098a17–18 & VI & X.6–8; see also Kraut 2018, ch. 10; Pakaluk 2005, 316–331.)

It is from this that Aristotle’s examination of actions and responsibility follows: As parents, teachers, legislators, fellow citizens and other possible authorities in a person’s life direct their actions toward virtues (via teaching as well as via disciplines and rules, and related honors and punishments), the authorities would presumably want to legislate the honors and punishments – or *rewards* and punishments, or *praise and blame* – appropriately. It stands to reason, then, that the honors and punishments should be sensitive to the character of the person who performed the action. Thus, for the person to be judged appropriately, it is important to find out when, and to what extent, actions are signs of an underlying character. The question is: When ought someone be considered responsible for a particular act they performed, so that we can be confident that blame is directed at someone whose action originated from their (bad) moral character and not something external? (Pakaluk 2005, 118–119.)

34. Notice that “to be compelled” and “compulsory” – i.e., “compulsion” – should be understood very specifically here. It is not used in its vernacular meaning of “an irresistible urge or desire to behave in a certain way”, but with the meaning of some outside force intervening and forcing or coercing the action. This applies throughout the thesis, when the concept is used, though in later chapters with the added recognition that also internal things like brain tumors or mental illnesses can be thought as external interventions to the agent’s cognitive system or character, leading to involuntary acts.

35. Aristotle briefly mentions that also *voluntary feelings* are generally praised and blamed (he once mentions “[voluntary] feelings and actions”). However, afterwards he only focuses on actions. It may be interpreted that he views feelings and actions to be connected in such a way that when actions are discussed feelings are covered as well. As Reeve writes, Aristotle may view “voluntariness or involuntariness of feelings to be parasitic on that of actions” (see III.1.1109b30–35, n198). In a similar(ish) vein, Eshleman (2016, ch. 1) describes Aristotle’s reference to “feelings” to refer to dispositional traits of character. However, the translation on the part of “feelings” has also been suggested to be rejected (Pakaluk 2005, 123n2). In any case, I will follow Aristotle and likewise focus on actions, while interpreting that feelings and actions (or character traits and actions) may be intimately connected.

36. In the context of Aristotle’s examination of actions, I would interpret his expression “pain” to refer to negative psychological experiences for the agent due to having performed an act, and “regret” to mean that the agent clearly indicates that they would have acted differently had they not been ignorant of some relevant knowledge. Where pain is felt, it would also seem to implicate the agent to wish they would have known what they didn’t know so they could have acted differently. It is useful to note that the Greek terms for “pleasure” (*hêdonê*) and “pain” (*lupê*) have a broad meaning, covering also “liking” and “disliking” as well as “satisfaction” and “dissatisfaction” (Pakaluk 2005, 103). Pakaluk (2005, 125) also uses the fitting term “distress”, and Campos (2013, 105) uses “repentance”. I will use the expression “pain and regret” in this section due to these terms being used in Reeve’s translation (Ross uses “pain and repentance”, and Rackham uses both “pain and regret” and “sorrow and regret” at different parts). However, later on (e.g., in sect. 6.1.1.4), I will use the less ambiguous “other-regarding sorrow and regret” to express a similar qualifier. (see III.1.1110b15–19 & III.1.1111a16–21.)

37. “Moral agents” include people (and possibly other similar enough creatures) who have the capacity to make moral judgments based on *some* conception of good and bad, and who are thus expected to meet the demands of morality (insofar as requirements for moral responsibility are met in the case of a given action). Usually, moral agents are understood to not include, for example, young children, non-human animals, plants, or inanimate objects. (Haksar 1998.)

38. Remember that if pain and regret is not felt in the case of ignorance [B], the action is *non*-voluntary and thus liable to responsibility. Arguably, pain and regret would also be required for a compulsive act [A] to qualify as involuntary, after the compulsive force has ceased to exert its power onto the agent (Campos 2013, 109–110). However, this seems to be a pointless addition because in practice it is irrelevant: an involuntary act due to external force is one to which an agent “contributes nothing” (III.1.1110b1–18; Pakaluk 2005, 123–124). Insofar as the act was something they would not have opposed to, they would have typically contributed something, and hence the act would by definition not have been forced (Pakaluk 2005, 123–124). Still, by the same token, an agent would feel *pain* for a forced act as they would oppose to it (III.1.1110b10–12). However, unlike Campos (2013, 109–110), Aristotle does not mention that the agent would be expected to feel *regret* for a forced act. This seems reasonable, because regret can seem pointless for an act that was not under one’s own control. Regret would seem appropriate only for the one who did the forcing (if it was a person), as the act was practically in their control.

39. As it happens, Aristotle considers children and (non-human) animals voluntary agents, yet not in the same way responsible as adults would be – making them an exception in terms of people being held responsible for voluntary acts. This seems to be because even though children can perform voluntary actions, they are incontinent (i.e., they display ‘lack of mastery’; *akrasia*) and cannot perform *deliberate choices* (requiring sufficient reasoning capability to make evaluations between choices). Aristotle views deliberate choices to be better indicators of character than voluntary actions. (III.1.1111a21–2.1112b17; see also Pakaluk 2005, 129–140.) For example, perhaps it might be thought that if I merely voluntarily decide to throw a rock at someone – say, due to momentary frustration – it may only tell something of my intuitively acting character. But if I thoroughly deliberate *a priori*, and still decide to throw the rock, that tells something of my deliberately acting character. The latter character seems more sinister of the two.

40. By examining other Aristotle references and by process of elimination, Reeve notes that the original Greek [3] *peri ti* and [4] *en tini* can together be understood as [4] expressing the somewhat vague [3] more precisely; and what Aristotle

otherwise refers to as *the one being affected* (ancient Greek: *on*). Reeve suggests this can be understood as explicating that the one affected by the action is the one in which or in whom the action occurs or is located. Thus, Reeve states, the location of an action is one of the particular factors in which the action lies. The other translations make this distinction in a varying degree, both only explicitly listing six particulars of the situation. Ross' translation seems most limited, as he lists on this part only "what or whom" the agent is acting on, when combining 3 and 4 into one. Rackham is substantially clearer, as he translates the list referring to the "nature and number of circumstances" of which the agent can be ignorant of, and that are: "[1] the agent, [2] the act, [3] the thing that is affected by or is the sphere of the act; and sometimes also [4] the instrument, for instance a tool with which the act is done, [5] the effect, for instance, saving a man's life, and [6] the manner, for instance, gently or violently". (III.1.1111a2–6.)

**41.** In the contemporary debate about the epistemic condition, 'culpability' and 'blameworthiness' commonly have the same meaning (Wieland 2016, 3n3). In the context of this thesis, these are also used synonymously, unless otherwise indicated.

**42.** This remark foreshadows chapters 6 and 7, and any major disagreements that can be interpreted between Aristotle and me. I will not, however, analyze any contemporary accounts of the epistemic condition in terms of how they would strictly compare to Aristotle's. I make this omission because of a combination of the following reasons: (i) there is room for interpretation in Aristotle's account, and (ii) he is often merely referred to as the first to make the relevant distinctions, and hence (iii) none of the contemporary discussion revolve around him in the sense that he would be meticulously cited (in fact, any sort of cited outline even roughly approaching the one presented here seems absent from all recent discussions about the epistemic condition). Thus, I leave it for the reader to deduce, between the lines, the contemporary connections to Aristotle. Suffice it to say that different contemporary accounts of the epistemic condition may be read to emphasize the elements of Aristotle's original conceptualization in various, often mutually conflicting ways (see sect. 5.4 & 6.7).

**43.** The remaining chapters in Book III start to stray from our primary interests. However, it is illustrative to briefly outline them as Aristotle does discuss some further nuances of the themes laid out in the key chapter 1. Namely, in chapter 2 he explores the nature of *deliberate choice* and how it seems a better discerner of people's character than their actions. He characterizes it as something not to be confused with appetite, spirit, wish, or belief. By comparing deliberate choice to these things that do not characterize it, he ends up describing it as something that is voluntary and involving reason and thought utilized to make evaluations between choices. Notably, not everything voluntary is deliberately chosen (for example, instinctive actions). In chapters 3 and 4, Aristotle continues to explore deliberation and wish. For example, he views that we (primarily) deliberate about things that are up to us and doable in action, and about means but not ends; and that exercising virtue is concerned with what we can deliberate and choose as reason, or practical wisdom, directs. In chapter 5, Aristotle describes how he views our character, along with virtue and vice, to be – in some sense – up to us; within our capability to choose. (Pakaluk 2005, 143–149; cf. Eshleman 2016, note 7; Talbert 2016, 76–79).

In the rest of Book III and the rest of the *Nicomachean Ethics*, Aristotle focuses on discussing, for example, specific character-related and thinking-related virtues; vices as well as lack of self-control (*akrasia*) and beastliness that are all to some degree something to avoid in character; pleasure; friendship that, in certain forms, can support virtue if not be one itself; and the importance of politics, setting the stage for his further work *Politics*. Overall, Aristotle lists approximately 10–14 virtues, with somewhat varying lists – with also differing terminology – provided by, for example, Aristotle & Reeve (2014, 239n185), Flanagan (2011, 95), and Pakaluk (2005, 2–3). The four classical "cardinal virtues", on which all other virtues hinge, are arguably self-mastery (i.e., temperance), courage, justice, and prudence (i.e., practical wisdom) (Pakaluk 2005, 167–168 & 167n12 & 185–186). In any case, Aristotle's list does not seem to be very fixed, but more of an outline. Thus, I think Aristotle's virtue ethics ought to be thought as a part of a process of searching for virtues, and aiming to cultivate them, rather than as a final say on the matter. A more widespread cross-cultural search seems to be an emerging trend in the field of virtue ethics (Hursthouse & Pettigrove 2018, ch. 4; see also sect. 6.6.2–6.6.3).

**44.** Unlike Sher, philosophers Michael J. Zimmerman (2017, 219) and Fernando Rudy-Hiller (2018) emphasize that these two conditions are also usually taken to be jointly sufficient conditions for moral responsibility. Both do agree with Sher, however, that the epistemic condition has been largely neglected until very recently – compared to the freedom requirement.

**45.** Sher mentions that philosophers John Martin Fischer and Mark Ravizza have termed the first condition "freedom-relative condition", and the second "cognitive condition" (Sher 2009, 3n3). Philosopher Peter A. Graham has also called the freedom requirement "the metaphysical condition" (e.g., in Robichaud & Wieland 2017, ch. 9).

46. Discussion of the freedom requirement can be found, for example, within the large body of literature produced by compatibilists and incompatibilists discussing determinism and free will (see, e.g., McKenna & Coates 2018; Vihvelin 2017; see also sect 2.2).

## Section 2.2

47. Aristotle argued against a version of *fatalism* but seems not to have recognized causal determinism (e.g., in his work *On Interpretation*, ch. 9; see also Aristotle & Reeve 2014, III.5; Pakaluk 2005, 143–149). The difference is important, however. Fatalism is a more radical view: everything happens in a set way, *regardless of what we do*. (Eshleman 2016, ch. 1; Talbert 2016, 16.)

It is important to notice that even if the universe is causally deterministic – including our thoughts and actions – that does not mean our thoughts and actions do not matter. It makes a crucial difference whether we choose to, for example, lay on our bed for the rest of our life or to perform rational actions in accordance with virtue, even if the choice we experience making is ultimately determined. Similarly, for example, arguments and deliberation, and the distinction between voluntary and involuntary actions, would be important parts of the deterministic system and our experience within it and as parts of it. (ibid.)

48. Before the more scientific modern era, *theological determinism* was especially influential, at least in Europe. It identifies the relevant antecedent conditions as the nature and will of (some) God or gods. (Eshleman 2016, ch. 1.)

49. Incompatibilism and compatibilism can be connected to the ancient Greek schools of philosophy of Epicureanism and Stoicism, respectively (Eshleman 2016, ch. 1).

Nowadays, after the apparently indeterministic findings in quantum physics, it may well be that on some level determinism is false. However, this does not seem to affect the views very much as prior states of the universe combined with the (in some part indeterministic) laws of nature would appear to still suffice for the relevant antecedent conditions. Incompatibilists can still maintain that indeterminism and determinism are *both* just as much outside the requirements for free will, and compatibilists can likewise maintain their position of understanding free will in some specific way that makes it compatible with any indeterminism *or* determinism there may be (see, e.g., S. Harris 2012, 2014; cf. Dennett 2014; see also Caruso & Dennett 2018). The kind of (incompatibilist) free will skepticism in question here goes more precisely by the name *hard incompatibilism*, distinguished from the pre-quantum-era *hard determinism* (Caruso 2018, sect. 2.1–2.2). In the body text, whenever I use the term “determinism”, it can be understood to include any indeterminism there may be in the laws of nature.

50. Quite interestingly, approximately 125 years after having been hypothesized (i.e., argued) by philosopher Friedrich Nietzsche, it has been suggested via multinational empirical support that, at least in part, punitive desires – associated with the merit-based view – *motivate* belief in “free will” (largely based on individuals’ own intuitive definitions). Arguably, it is easiest to hold people responsible when we attribute, in some sense, free action to them. Thus, if we are motivated to hold someone responsible, the easiest first step may be to accept free will (at least unconsciously). (Clark et al. 2014; see also Clark, Winegard, & Baumeister 2019; Feltz & Millan 2015; Martin, Rigoni, & Vohs 2017.)

In recent years, there has been interest in experimental philosophy on whether people are intuitive compatibilists or incompatibilists (so that we could know what view has more argumentative burden, and what it is we may be revising). The debate is likely far from settled, and I encourage visiting the articles mentioned in this footnote to get a fuller view, but it does appear that people are, in a sense, both. Folk intuitions seem to be messy and incoherent, especially as they appear to be guided by situationally varying motivational factors (Clark et al. 2019). Moreover, even philosophers may be affected, with compatibilist judgments being especially vulnerable to motivated reasoning to avoid a potential threat to moral responsibility (Clark et al. 2019; see also Clark et al. 2014.).

One thing seems clear, however: being exposed to theory can affect folk intuitions, while also affecting behavior relating to responsibility. For example, learning about anti-free-will arguments or neural bases of human behavior can reduce belief in free will, while also reducing support for retributive punishment (Shariff et al. 2014) and increasing support for rewarding positive behavior (Genschow, Rigoni, & Brass 2017; for a summary of some further associations between belief in free will and behavioral/psychological outcomes, see Chandrashekar et al. 2019).

51. There is no strict agreement on who belongs to what category, though. For example, Sher (2009) and Zimmerman (2010) have categorized Thomas Scanlon and Angela Smith as attributionists, while, for example, Eshleman (2016, sect. 2.2) distinguishes them to represent the more specific unifying category of answerability theorists. Eshleman’s categorization, however, draws from a relevant paper by A. M. Smith (2012) that was published after the earlier

categorizations. Thus, it seems to reflect a distinction that emerged after Sher's 2009 book. Furthermore, not all agree that answerability provides a unifying view, rather than being a distinct category itself (see Shoemaker 2011). (see also Wieland 2017, 4n6.)

## Chapter 3

### Section 3.1

52. This roughly corresponds with what Michael J. Zimmerman calls the *unqualified* searchlight view (Zimmerman 2009). In section 5.1, I introduce Zimmerman's *qualified* searchlight view. Although Sher does briefly note and consequently rejects the qualified searchlight view (Sher 2009, 33–39) – albeit not using the term – he clearly focuses on the unqualified version in his examinations. Therefore, I save the examination of the qualified thesis to section 5.1 via Zimmerman.

### Section 3.2

53. This refers to Zimmerman's *origination thesis*, which I examine more in section 5.1.2 (see also sect. 5.4.3).

### Section 3.3

54. The same would apply to prudential responsibility as well, that Sher is keen to add into the mix. Since I am interested primarily in moral responsibility, I will not explicitly examine prudential responsibility here. However, it should be noted that everything Sher says can be applied, with little adjustment, to prudential responsibility as well (Sher 2009, 29–31).

55. These cases are first introduced in Sher 2006b.

56. Sher is using “acts” in a broad sense here; encompassing omissions as well as positive actions (2009, 23).

57. Sher is firm on the cases 1–6 being ones where the agent would definitely be blamed and might well be liable to punishment, but a bit more cautious on cases 7–9, saying merely that in the latter cases the agents *seem* blameworthy and one (amerika) definitely deserving punishment. (Sher 2009, 24–28.)

### Section 3.4

58. This basic characterization of the distinction between practical and theoretical reason is sufficient in understanding what Sher is alluding to. However, there are other interpretations about the distinction as well (see Wallace 2018, ch. 1).

59. The arguments Sher examines are presented in Bok 1998, and Korsgaard 1996a.

60. Sher has discussed some of these issues in more detail in chapters 2 and 3 of his earlier book *In Praise of Blame* (2006).

61. In broad terms, Sher describes the principle being connected with Kant's theory of moral worth, even though it has no particular connection to Kant's theory of the right (Sher 2009, 56). Sher recommends Williams (1985) as a further source for a statement of the view that Kant's view of moral worth reflects an ideal of fairness.

62. Sher states that the contemporary philosopher R. Jay Wallace, a compatibilist, seems to have paid the most attention to the Kantian Principle, for example in his book *Responsibility and the Moral Sentiments* (1994). However, Sher sees that even Wallace has seemed more interested in the meaning and implications of the principle rather than in the question of why, if at all, the principle should be accepted. (Sher 2009, 57.)

63. Sher considers it given that attributions of responsibility cannot be based on facts about any individuals other than the relevant agents themselves, for it would be incoherent “[b]ecause responsibility is widely viewed as a prerequisite for the

legitimacy of praise, blame, and punishment, and because praise, blame, and punishment are by their nature reactions to particular persons” (Sher 2009, 57).

64. Sher himself acknowledges that his conclusion threatens the control condition; i.e., the idea that responsibility requires control, albeit threatens it only if control is understood in the sense that he is inclined to understand it: as requiring awareness of one’s options. (Sher 2009, 144–146.)

## Chapter 4

### Section 4.1

65. This formulation follows after Sher having initially examined and dismissed various versions, or interpretations, of the common-sense appeal that an agent “*should* have known better” (Sher 2009, 71–84). As I more explicitly summarize in section 5.1 via examining Zimmerman’s Thesis C, that Zimmerman calls the “knew or should have known” thesis, perhaps the most pertinent problem Sher (and Zimmerman) see in those views is that they fail to account for genuine situational excuses. Here, I focus on Sher’s own views that he sees not to fail. He calls his approach to the idea that an agent “should have known” as indirect (see clause 2a+2b), whereas the views he dismisses are more so direct and what he considers problematic versions (see 2a alone) (Sher 2009, 75 & 78 & 85–86). Specifically, Sher considers his version less problematic because it doesn’t merely focus on the agent’s failure of recognition by itself but takes account of psychological features of the agent, thus establishing what he considers a justified *causal link* between the agent and the act’s wrongness (or foolishness) while also enabling us to account for genuine, responsibility-redeeming excuses (Sher 2009, 86–88; see clause 2).

66. ‘Wrongness’ of an act here implicates (negative) moral responsibility, and ‘foolishness’ of an act implicates (negative) prudential responsibility. As the interest of my examination lies in moral responsibility, I will not explicitly illustrate nor address the prudential aspects of Sher’s arguments. However, as notated before, everything Sher says can be applied, with little adjustments, to prudential responsibility as well. PEC and FEC are intended, on Sher’s part, to apply to both moral and prudential responsibility.

67. Sher has defended the claim that lack of moral flaws can lead to immoral acts, with some illustrations, in chapters 2 and 3 of his book *In Praise of Blame* (2006a). For example, an agent who always thoroughly and even obsessively aims to be generous and empathetic may encounter situations where their relentlessness results in working against their moral aims via an explosive psychological over-encumbrance (see Sher 2006a, 23–24). Via discussing illustrations of these kind, Sher argues against a view that Scottish philosopher David Hume, among others, have endorsed: against a view that agents are only blameworthy for acts that reflect flaws in their character (Sher 2006a, ch. 2–3; cf. Hume 1739–40, T 2.3.2.6, SBN 410-1). The larger argument in the book – also largely relying on Sher’s intuitions – is that blame is *conceptually* inseparable from our commitment to morality and characterized as being something a person can be worthy (see Sher 2006a, 14–16 & 132-133). However, unlike in *Who Knew? Responsibility Without Awareness* (2009), in *In Praise of Blame* (2006a) Sher does not argue which individuals are blameworthy (see Sher 2006a, 133–135).

68. Specifically, Sher notes that “it is not immediately clear [in the homeopath case] whether we should say ... that his failure to recognize his act as wrong is substandard for someone who has easy access to orthodox medical care and the Internet, or that that failure meets the standards that apply to those who have easy access to orthodox medical care and the Internet but who also believe strongly in homeopathy.” (Sher 2009, 98.)

69. The reasonable-person standard, in legal literature, is a standard used to determine “who is criminally or civilly negligent, who has legitimately acted in self-defense, and who is legally liable in many other contexts” (Sher 2009, 100). Generally, the standard can roughly be characterized as viewing a negligent agent as someone who should be aware of the risky nature of their conduct with a reference to the care that a reasonable person *could* and *would* exercise in similar situational circumstances (see Sher 2009, 80). In legal literature, there is much debate concerning the boundaries of the standard; in terms of where and how to draw the examined line between the agent and their situation (see Sher 2009, 100–104).

70. This claim seems quite reasonable, even though the most often considered aspects of personhood in philosophy – which seem to be *psychological continuity* and *psychological connectedness* – are only implicitly included. At the same

time, it seems Sher considers the also often considered physical aspects of personal identity to be largely irrelevant in this context. (see Olson 2017.)

71. Sher's metaethical position *seems* to be some form of moral realism, specifically *intuitionism*. He has expressed in an essay *I Could Be Wrong* (2001) that being "clear-eyed moral agents" requires us to understand that we could be wrong in our moral judgments due to them being always contingent on our individual upbringing. Other possible answers to the challenge of our varied backgrounds affecting our moral judgments, that Sher sees, would be either moral skepticism or re-evaluating evidence for our moral views to further justify them, but he supports simply admitting we could be wrong while continuing to make and deliberate our moral judgments (Sher 2001).

72. A more thorough and recent treatment of Levy's position can be found in his book *Consciousness and Moral Responsibility* (2014), in which he states that he would *not* describe himself strictly as a volitionist, at least not in the minimalist sense that Sher is using the label (127–128n8).

73. The key sources for these views are Scanlon 1998, and A. M. Smith 2004, 2005, 2008. See also A. M. Smith 2012.

74. Angela Smith (2010) seems to generally accept this characterization, though she does have some critical comments that I introduce in section 5.2. Insofar as her and Sher disagree, she sees the matter as a family dispute.

75. Sher notes that the name "attributionism" derives from this view: an agent's responsibility is restricted to the attitudes and actions that reflect their evaluative judgments, and in this sense *attributing* responsibility to the agent. Philosopher Gary Watson (1996) has distinguished responsibility as attributability and responsibility as accountability, the first of which Sher mentions to be a relative of Scanlon and Smith's view. (Sher 2009, 129; see also sect. 2.2.)

76. Sher describes the idea being first introduced in Frankfurt's essay *Freedom of the Will and the Concept of a Person* and explored in other papers in his collection *The Importance of What We Care About* (1988). The idea has been further developed, in two different ways, by Gary Watson (1975), and Christine Korsgaard (1996b, lecture 3).

77. This view is extrapolated by, for example, John Martin Fischer and Mark Ravizza (1998, ch. 7), and Alfred Mele (1995).

78. Further criticism of this group of views has been presented by, for example, Nomy Arpaly and Timothy Schroeder (1999), and Susan Wolf (1990, ch. 2).

79. This sort of view has been expressed by Susan Wolf (1988).

80. The example about Daniel can be considered to concern *prudential* responsibility, but I think it best illustrates Sher's distinction between responsible and non-responsible agents.

81. Based on his Humean social intuitionist model (SIM), social psychologist Jonathan Haidt has developed a fitting rider-and-elephant metaphor to illustrate what he sees to be the cognitive processes behind our moral judgments. This, or something alike, may be what Waller is alluding to. Specifically, the metaphor illustrates the relationship between the initial cognitive step of intuitive, *a priori* judgments, and the step of *post hoc* reasoning of and via those judgments. In the metaphor, the elephant represents our automatic evaluations and judgments – while the rider, who is riding the elephant, represents our conscious reasoning and rationalizations. As the rider and the elephant encounter moral events in the world, the elephant automatically moves to the direction of any moral judgments it makes, while, after the fact, the rider who is the elephant's spokesperson can only try and rationalize why the elephant moved to the direction it did. In other words: reason (the rider) is the servant of the intuitions (the elephant) (though not the "slave" of the "passions" or "sentiments", as Hume thought). Still, social reasoning and deliberation can change future intuitions, as can plain (in-)group pressure. Private reflection can also work, albeit it is utilized much more rarely than philosophers would perhaps like to think. Nevertheless, reason always remains the servant of intuitions, under the social intuitionist model of moral judgment. (Haidt 2001; Haidt 2012, 52–56 & 61–83; see also Haidt 2011b; Hume 1739–40, T 2.3.3.4, SBN 414-5; for a review of models of moral reasoning, including SIM, see Guglielmo 2015; for an introduction of dual process theories, see sect. 6.1.3.)

## Section 4.2

82. Once again, Sher explicitly includes prudence's nature in his examination as well, in addition to morality. Albeit I am focusing on morality, everything Sher says here he considers applying to prudence as well.

83. All other Sher's examples are from the prudential domain but are illustrative enough to being mentioned: "the successful investor who 'instinctively' purchases some stocks while avoiding others, the experienced diagnostician who recognizes his patient's condition despite the absence of clear symptoms, and the sailor who senses the onset of a storm while the sky is still a brilliant blue." (Sher 2009, 142.)

## Section 4.3

84. In its classic form, the account lists only external factors (see Hobbes & Curley 1994, I.XIV.79). A more recent variant extends the approach to include also some internal ones (see Ayer 1997).

85. See Hume 1739–40, T 2.3.2.6, SBN 410-1 – T 2.3.2.7, SBN 411-2. See also Hume 1748/1777, sect. VII, part II.

86. There are many variants for this approach. For example, see MacIntyre 1957, Nozick 1981, and Dennett 1984.

87. See J. M. Fischer and Ravizza 1998.

## Chapter 5

### Section 5.1

88. These statements, that seem to boil down to Sher's intuitions about the example cases, are problematized in chapter 6, where – among other things – I emphasize the problems of relying on intuition, the importance of noting metacognitive knowledge, and the value of taking note of the deliberative perspective.

89. A more in-depth treatment can be found in Zimmerman 2008. Furthermore, an alternative, demonstrative reconstruction of the regress that the argument suggests has been outlined by Rudy-Hiller (2018, ch. 2) as well as Wieland (2017, 12) (see also Talbert 2016, 130–135; sect. 5.4).

90. In this context "wrong" means overall morally wrong (noting the consequences), not merely *prima facie* morally wrong (Zimmerman 2009, 256). This footnote is my own addition; not included in the quote.

91. Zimmerman adds that in his thesis culpability is rooted in a *belief* about wrongdoing, and that it seems the required belief is an *occurrent/conscious* one (rather than merely dispositional), with one possible exception. Namely, the belief may not need be occurrent in the case of routine or habitual actions as a belief may not play any part in such actions. That is, Zimmerman views that an agent can be culpable only for actions if their (conscious or unconscious) belief concerning wrongdoing plays a role in the reason why they perform the action. Thus, Zimmerman's view is not strictly the qualified searchlight view, as presented above, as the searchlight view would only focus on the agent's occurrent/conscious awareness, without exception. (Zimmerman 2009, 258; see also sect. 5.4.1.2n103.)

92. As I underline in section 6.1 (albeit in a different way than Zimmerman), I think clause 2b, in Thesis B, can be understood as a specifier: sometimes the act or omission we seem prone to deem the most relevant, when hastily judging someone's moral responsibility, may not in fact be the most relevant one or one that we should pay most attention to (cf. the relevancy of Alessandra, in Sher's example case 1, being (un)aware of her dog at a given moment vs. the relevancy of Alessandra being (un)aware of the metacognitive or preparatory nuances that keeping her dog safe would pertain before a given moment, for example, in unpredictable situations occurring while leaving her dog in a van).

93. It appears that Zimmerman's premise (4), with its requirement of "control", refers to a *compatibilist* variation of free will (i.e., a variety where determinism and free will are seen compatible). To make this more explicit, one might imagine Zimmerman requiring "freedom" of the agent, instead of "control".

94. Zimmerman has further argued that punishment is immoral and thus wrong (2011).

## Section 5.2

95. Smith herself does not call her view attributionism but “rational relations view”, referring to the rational relation “between what we notice and what we evaluate or judge to be important or significant” (A. M. Smith 2005, 242; see also Talbert 2011, 149). Later, Smith has referred to her view as *answerability* (A. M. Smith 2012; see also sect. 2.2).

96. Similar to Smith’s reply, Matthew Talbert (2017) also argues that the connection cannot be merely causal but that it also needs to be grounded in a *lack of moral concern*.

## Section 5.3

97. Waller’s example does not seem to contain a situation that would involve *moral* or even necessarily *prudential* responsibility. Nor does it handle “constitutive elements” of an agent in the same way as Sher: to refer to the causal structure that sustains the agent’s reason-responsiveness. Still, the example demonstrates the nature of the problem well: it may be that in some situations it is not reasonable to expect an agent to perform well cognitively, much like in some situations it is not reasonable to expect them to perform well physically. In such cases, the agent may be reasonably excused from responsibility. This is reminiscent of Smith’s basic objection to Sher’s account (see sect. 5.2).

98. “Ego depletion” refers to the psychological concept named by Baumeister that self-control or willpower draws from a limited pool of mental resources that can be used up. The phenomenon has been supported in many studies and in an independent meta-analysis over the last 20 years, but nevertheless its existence as a phenomenon has recently been challenged due to alleged publication bias and other shortcomings. Suffice it to say that currently the idea seems disputed. (Cunningham & Baumeister 2016.) Nevertheless, Waller’s larger point seems to hold: situational factors that affect the agent’s constitutive cognitive functions – whether via ego depletion, fatigue, stress, or some other similar psychological phenomenon – may compromise them to errors that may be viewed to render the agent pardonable in a sense analogous to the sprinter.

99. Waller’s own stance (2011), very roughly, seems to be that moral responsibility – that he distinguishes from take-charge responsibility that concerns taking charge of one’s own life – is a failed concept that is fundamentally unfair and harmful, and its thorough abandonment would be profoundly beneficial for everyone.

100. Nelkin’s own compatibilist interpretation of moral responsibility seems to rely heavily on examining the freedom requirement (2011a).

101. Talbert’s own views seem to be close to, though not completely in line with, those of Thomas Scanlon and Angela Smith (see Rudy-Hiller 2018, sect. 3.3; Talbert 2011; Talbert 2016, 145–152; also see section 5.2 & 5.4). He examines the question of moral responsibility in his more descriptive rather than prescriptive overview of the field *Moral Responsibility*, (2016), Polity Press.

## Section 5.4

102. Compare this to what a central question might be if the consequentialist view was more prevalent: *When* – in a consequentialist sense – is it reasonable to consider (and hold) agents responsible, so that we may expect them to henceforth do better, and for our responsibility impulse to not backfire? (cf. sect. 2.2 & ch. 6 & 7; Wieland 2017, 5n7.)

103. Some think *true belief* is required (e.g., Peels 2014, 493; Rosen 2008, 596–597). Some think *reasonable* or *justified belief* is required (e.g., Ginet 2000, 270). And some think that only *belief* is required (e.g., Zimmerman 2008, 198), because – as the argument goes – it demonstrates a particular kind of culpability that one incurs when one knowingly defies what one takes to be the requirements of morality, regardless of whether one gets them right (Levy 2011, 142; Zimmerman 2017b, 91). (Rudy-Hiller 2018, sect. 1.2.)

**104.** Quality-of-will theorists tend to agree that *de dicto* awareness is not required for moral responsibility, but do not agree whether *de re* awareness is required (see Rudy-Hiller 2018, sect. 3.3; see also sect. 5.4.1.1).

## Chapter 6

### Section 6.1

**105.** As we are talking about intuitions, it seems noteworthy that regardless of some possible similarities, I originally formulated my considerations before being aware of any replies to Sher's account.

**106.** Of course, she may be commonly held responsible and even in fact be legally responsible, but that is not what is at question here nor in the other cases. The question is, should we suspect the universality of Sher's intuitions in the case of *moral responsibility*.

**107.** "Metacognition" refers to the ability and skill to reflect on our cognitive processes; i.e., thinking about thinking. It is relevant for the questioning and re-evaluating of one's knowledge as well as for regulating one's cognitive performance. The concept is often used in the context of education and bettering learning performance via awareness of one's learning-related beliefs and behavior. However, it is applicable to any context that involves complex information that should be effectively processed. (see Flavell 1979; Proust & Fortier 2018; Schraw 1998; see also sect. 6.4.2.2.)

Metacognition is traditionally classified into two components: metacognitive knowledge (or awareness), and metacognitive regulation (or control). *Metacognitive knowledge* refers to an agent's cognitive monitoring of themselves as processors of information that are divided into three types: declarative knowledge (knowledge *about* what factors can influence one's performance), procedural knowledge (knowledge *how* heuristics and strategies can enable one to perform better, even automatically), and conditional knowledge (knowledge *why* and *when* to utilize the former two, and to allocate cognitive resources accordingly and use strategies more effectively). *Metacognitive regulation* refers to regulation of cognitive activities in ways to better one's learning and performance, further divided into three skills: planning (selection of strategies and allocation of resources so that a given task is better performed), monitoring (awareness of one's comprehension and task performance), and self-evaluating (appraising one's performance *a posteriori*, including re-evaluation of the strategies that were utilized). Some theorists have added further components to metacognition as well, for example *metacognitive experiences*, referring to experiences that are relevant for an ongoing cognitive task (e.g., for taking care of a dog). (Flavell 1979; Schraw 1998; see also Proust & Fortier 2018.)

Thus, more specifically, what Alessandra would seem to gain is metacognitive knowledge, which enables better metacognitive regulation in similar situations in the future, thus more reliably producing metacognitive experiences relevant to taking care of Sheba (e.g., by following pre-emptive heuristics and strategies or just by being better passively aware of one's performance). This can also include enhanced *metamemory* (i.e., monitoring and control of one's memory performance), *mnemonic strategies* (e.g., enabling appropriate memory functions to trigger when stepping out of a car with a dog inside on a hot day), and *meta-awareness* (i.e., awareness of mind wandering) (Seli et al. 2017; Tauber & Dunlosky 2016). Metacognitive performance gains seem to have a lot to do with enhanced pattern recognition via acquiring information about one's pattern recognition (e.g., acquiring information about one's biases and concrete cognitive errors like in the failure of keeping track of a memory about a dog in a car).

I will be using the shorthands "metacognition", "metacognitive knowledge", and "(meta)knowledge" (referring to knowledge that may include knowledge, metacognitive knowledge, or both) to refer to these kinds of various cognitive processes that enable a person to perform better in the future, and essentially to better Socratically 'know thyself'. This betterment can be understood as betterment in terms of both active and passive knowledge and awareness.

**108.** It could also be added that shooting *even* a burglar (let alone an assumed burglar) could be considered a blameworthy decision. However, sorting that out would require its own moral argument (and which would require awareness of itself from anyone we would see acting against its conclusion, before my intuitions would deem relevant moral responsibility justified).

**109.** Good humor seems to be a particularly risky mix of violating a perceived norm in a benign manner. Mishaps would thus seem to be an expected occupational hazard, especially insofar as norms are subjectively perceived and locally evolving. (see McGraw & Warren 2010; McGraw, Warren, Williams, & Leonard 2012.)

**110.** The harm principle is British Enlightenment philosopher John Stuart Mill's principle that he initially formulated in his liberal defense of free speech in *On Liberty* (originally published 1859), and which sets the limitation of free expression

to where harm to others would be prevented. However, Mill was not very clear on what exactly he meant by “harm”, and there has been much debate on the matter since (see, e.g., van Mill 2018).

**111.** In Sher’s formulation of the case, Ryland is herself bewildered and a bit hurt by the reaction of the audience. Thus, it could be added that I do not think Ryland being hurt herself bears any weight on these issues, and I suspect Sher agrees. As an analogue: When summarizing Aristotle on the Internet, I may be bewildered and a bit hurt if someone gets offended by some aspect of the summary, but my reaction bears no weight on what we should think about summarizing Aristotle on the Internet in general (nor does merely someone being offended bear any weight on the matter).

**112.** Bloom (2016) distinguishes affective/emotional empathy from cognitive empathy and compassion. By “emotional empathy” he refers to what Enlightenment era Scottish philosopher Adam Smith called “sympathy”: placing ourselves in another’s situation; feeling with another, sharing their feelings (16–17). “Cognitive empathy”, on the other hand, refers to the capacity to understand what’s going on in other people’s mind; what gives them joy and pain, what they see as humiliating or ennobling (36). And “compassion” refers to caring for another; feeling *for* and not feeling with another (138). These concepts seem to be distinguishable also on the level of brain activity (136–142). (see also Bloom 2017; Klimecki, Ricard, & Singer 2013; Klimecki, Leiberg, Ricard, & Singer 2014.)

Further, Bloom (2016) argues that there are many downsides to emotional empathy as it can not only cloud our decision-making and moral judgments – for example, via biasing us to favor our immediate in-group and individuals over statistics, consequently further opening ourselves to tribalistic exploitation – but it can also exhaust us (which may lead to empathy avoidance; Cameron et al. 2019). The downsides get often overlooked, but compassion seems to avoid the downsides while retaining most of the good sides. Sylvain’s situation seems like a textbook example of what some of the downsides of emotional empathy can look like but that could be managed via compassion training and/or other fitting precautionary measures (see Klimecki et al. 2013, 2014). Of course, empathy does remain important in some respects – for example, in facilitating a special connection with our spouse and children – but Bloom encourages caution towards romanticizing it.

**113.** Aristotle’s original criterion was for the agents to feel “pain and regret” (see sect. 2.1.1n36), but nowadays “sorrow and regret” seems less ambiguously fitting. Also, the extra qualifier “other-regarding” seems useful, to rule out the possibility of the agent being sorrowful and regretful *only* egocentrically (in cases like Sher’s examples, where the act primarily concerns others). Egocentric regret could be motivated, for example, by an upcoming, unexpected prison sentence, rather than actual concerns for those whom the regretted action negatively concerned.

**114.** It seems there might also be some overruling principles here. For example, if an agent has a neurological or psychological condition that hides or blocks his or her sorrow, but they still feel sincere other-regarding regret, such an agent might also be considered pardonable. In such a case their ability to express sorrow would be out of their control. Hence, the emphasis of the qualifier concerns (other-regarding) *regret*.

**115.** Better domain-general metacognitive skills may possibly, to some degree, be developed *before* encountering domain-specific situations where they would come in handy. However, the state of evidence seems either inconclusive or just emerging, and neither the teaching subjects nor possible methods that might enable this seem widely pervasive. (see Carpenter et al. 2019; see also Callender, Franco-Watkins, & Roberts 2016; Schraw 1998; Stokes 2012.)

**116.** It should be noted that hypothetical scenarios have been shown to not necessarily attain any intuitions we might have in corresponding real-life scenarios, and experiments of our judgments under hypotheticals – e.g., in dictator games – may drastically differ from experiments of our judgments in real life scenarios (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan 2014b). Nevertheless, our intuitions appear to differ. For a concise table summary of how Sher’s intuitions seem to differ from my alternative more charitable intuitions – which may be called the *metacognitive interpretation* – about the original cases, see [Appendix 1](#).

**117.** This is reminiscent of Aristotle’s idea that habit cultivates virtue, and virtue is often learned through practice (see Aristotle & Reeve 2014, II.1; see also sect. 2.1n33). After an agent having consolidated a virtuous habit, as an answer to metacognitive knowledge they have gained, it seems unlikely that they would still break the habit in the required ways. Especially when the habit, in this case, could merely refer to being better passively aware of what lead to their previous mistake, essentially requiring minimum conscious effort in similar situations, after the lesson has been accommodated into their metacognition.

**118.** It may well be that to some significant degree metacognitive performance is helplessly unreliable, even if at the same time improvements in average performance are possible. Hence my premise of a solidly learned metacognitive lesson, and hence no fault occurring in the future, would be hopelessly naïve outside the confines of a thought experiment.

One concrete possibility that might enable these sorts of lapses could be that metacognitive performance declines as we age. If this was the case, it would be more likely for any kind of (meta)memory mishap to happen the older we get. However, it seems the culturally commonsense assumption of memory performance declining merely as a function of age – as opposed to, for example, as a function of Alzheimer’s disease or decline of physical or social stimuli that are more common as we age – is often considerably overestimating the effect, and even getting it backwards in some important aspects of performance (e.g., adherence) (Fastame & Penna 2013; Hertzog 2003). In some contexts, internalizing this assumption may create a self-fulfilling prophecy of not properly attending to and learning from one’s errors (Hertzog 2003).

However, given the likely emotional baggage of the agents in the example cases, due to the highly morally salient nature of their errors, it seems likely that they would pay particularly close attention to their mishaps after they happen, regardless of their age (assuming an otherwise healthy cognitive function and moral understanding). *Insofar* as relevant cognitive performance does decline as a function of age (or via mental illness etc.), or otherwise remains universally unreliable, it would be something that theories of moral responsibility should take as seriously as the cognitive inadequacies in children.

**119.** The looming road I have in mind is well expressed, for example, by philosopher Galen Strawson in his article *The Impossibility of Ultimate Moral Responsibility* (1994/2013; see also Caruso 2018, sect. 2.3; Parks 2009). Literature concerning how *moral luck* implies restraints on responsibility is likewise relevant (see, e.g., Levy 2011; Caruso 2018, sect. 2.4), as are some scientific challenges for responsibility that have been raised (see Caruso 2018, sect. 2.5). To add, philosopher Derk Pereboom (2013) has provided an optimistic take on the skeptical possibility, especially in light of the phenomenon of moral outrage online (see also Caruso 2018, ch. 3; Pereboom & Caruso 2018). If all else fails, at some point these and similar views might be well worth revisiting. A summary of the overall discussion concerning moral responsibility skepticism can be found in a relatively new article in the *Stanford Encyclopedia of Philosophy* (see Caruso 2018).

**120.** Somewhat surprisingly, this is something that Sher might to some degree sympathize with. In the end, he may view that understanding and developing our own intuitions is merely the best we can do. (see Sher 2001.)

**121.** For some classic examples from the mid-20th century, many of which are still relevant today, see Gardner 1957. To gain some understanding of the underlying psychology behind the appeal of suspect epistemic claims, and of possible psychological requirements for thinking critically of them, see Lobato and Zimmerman 2018. For further cognitive perspectives, see articles by various writers in Pigliucci and Boudry 2013, part VI.

**122.** It is my understanding that research in both criminology and developmental psychology better support an intuition about the restorative, or rehabilitative, approach yielding better results, despite perhaps a more prevailing folk intuition that retribution would be the best approach (see Gershoff & Grogan-Kaylor 2016; Sherman & Strang 2007, 88; Sunstein 2018, 5). In any case, our conflicting intuitions do not resolve the matter, but research and relevant arguments do. And ideally our intuitions would, in time, adapt to the evidence. (see also Caruso 2018, sect. 3.4.)

**123.** For an example of a consensus statement on healthy eating from a group of researchers, see Oldways 2015.

**124.** The most obvious example seems to be the *double-slit experiment* in physics, first performed to light by British polymath Thomas Young in 1801, and to electrons by physicists Clinton Davisson and Lester Germer in 1927 (the Nobel Prize in Physics was awarded to Davisson in 1937). The experiment has later been extended to atoms and molecules (see, e.g., Eibenberger, Gerlich, Arndt, Mayor, & Tüxen 2013; see also Orzel 2013). Despite being a relatively straightforward physics experiment, it reveals and demonstrates the wave nature of light and matter – or, in popular lexicon, “wave-particle duality” of them – that continues to challenge many of our deepest intuitions. Yet, quantum mechanical phenomena seem to be relatively well understood via careful mathematics (although many would no doubt chuckle here, many physicists included).

Concerning the subject of thinking itself, Nobel Prize -winning behavioral economist and psychologist Daniel Kahneman, along with Amos Tversky and others, have made a long career in studying human thinking and biases. Consequently, the research has revealed many ways in which our fast and involuntary, intuition-generating thought patterns related to judgment and decision-making lead us astray, though we also couldn’t do without them. Kahneman has summarized much of his lifework in his book *Thinking, Fast and Slow* (2011).

**125.** *Cognitive dissonance theory* (Festinger 1957, 1962) suggests that when we confront novel information that contradicts our prior cognitions (i.e., knowledge, opinions, or beliefs), we experience a motivating mental discomfort – the state of cognitive dissonance – that calls for resolution. In other words, to effectively function in the world, we seek internal consistency in our ideas: that is, *consonance* contrary to dissonance. The more deeply held or subjectively valued any of our cognitions are, the more dissonance there is when contradicting information is encountered. The dissonance is commonly viewed to be reduced either via (a) changing one or several of the conflicting cognitions (e.g., changing one’s prior belief based on the novel information), or (b) adding new elements to one’s cognitions (e.g., coming up with justifications or rationalizations against the novel information), or (c) reducing the importance of the involved cognitions (e.g., ignoring or denying or suppressing/overriding the prior belief or the novel information). Often these processes involve social elements, like a friend or a group who assures us to one direction or another. (for a review of the theory, see Vaidis 2014; see also Harmon-Jones, Harmon-Jones, & Levy 2015; see also how the theory connects with fundamental attribution error [FAE] via Rosenberg & Wolfsfeld 1977; sect. 6.2.1n137.)

When encountering novel information, changing one’s prior contradictory cognitions appears to be more likely the *weaker and less emotionally* they are held. Conversely, coming up with justifications or rationalizations against novel information – or ignoring or denying it – even if the information is objectively supported, appears more likely the *stronger and more emotionally* the prior contradictory cognitions are held. For example, this can be seen when those who are more dogmatic politically are more prone to disconfirmation bias when encountering arguments contradictory to their position, and to confirmation bias when freely selecting sources (Taber & Lodge 2006). It has also been suggested that at least in some cases intuitive old beliefs are merely overridden, not overwritten, and can thus exist *simultaneously* with new more accurate unintuitive beliefs. This is indicated by a persistent lag in response time and comparative inaccuracy of subjects verifying between similar (a) accurate naïve and accurate scientific *intuitive* statements (e.g., “The moon revolves around the Earth”) versus (b) inaccurate naïve and accurate scientific *unintuitive* statements (e.g., “The Earth revolves around the sun”) (Shtulman & Valcarcel 2012; Shtulman & Harrington 2015). Moreover, two or more contradicting beliefs may be *compartmentalized* to avoid dissonance, meaning they are internally separated so that they may coexist, and are only referred to on two similarly distinct areas or contexts of life. (ibid.; for a review of cognitive dissonance reduction, see McGrath 2017; for a meta-analysis of selective exposure to information, see also Hart et al. 2009.)

**126.** These many cognitive phenomena derive from evolved mental processes we *all* utilize regularly, but it may be possible to attenuate their negative aspects via certain mind-training or debiasing techniques, or practiced ‘mindware’ (Nisbett 2016; Stanovich 2018a). However, it seems that ever since the 1970s there has been continuing disagreement regarding whether this is possible and if then to what degree (Yagoda 2018). I remain agnostic, though cautiously optimistic, on how well training or practice can induce permanent individual effects, but it seems undisputed that some of the negative aspects can be counteracted by institutional or social measures (e.g., via utilizing appropriate design ethics on social media platforms). Of course, they can also be exacerbated by unfavorable institutional or social measures.

**127.** Our impressions of what views are held within our perceived in-group, or in the wider society, can be mistaken just as our views themselves. The *false consensus effect* is a particularly interesting phenomenon, from the vantage point of living in the age of social media: we appear to be prone to relatively overestimate the degree to which others – especially our in-group (e.g., our “friends” on Facebook) – share our beliefs, judgments, and habits (for a review, see Marks & Miller 1987; see also Gilovich 1991, 112–122; Ross 1977; Ross, Greene, & House 1977). That is, our estimates of the commonness of a given belief, judgment, or habit tend to be positively correlated with those of our own (Gilovich 1991, 114). This may give us an illusory license to act with little inhibition (e.g., in our social media updates), and especially if *preference falsification* is taking place (see sect. 1.1n5). (see also fundamental attribution error: sect. 6.2.1n137.)

**128.** There are areas where our intuitions are useful, of course. For example, if we had a childhood free of neglect and abuse, we tend to be quite ‘sufficient’ in accurately interpreting others’ basic emotions via us unconsciously processing nonverbal cues. This can intuitively inform us of threats, for instance. Thus, these sorts of intuitions are usually worthy of listening. In fact, most of the mental processes in our moment-to-moment experience are intuitive, and intuitively learned, and we could not function without them (Kahneman 2011, 105). Also, felt emotions – that are most often intuitively interpreted – can be useful motivators. But if we trust our intuitions in the kinds of issues on focus here, concerning the evaluation of complex information in order to form accurate and trustworthy beliefs, we are highly prone to err.

These are some reasons why I talk about “taming”, not “eliminating”, our intuitive passions and sentiments. Not to mention that elimination would be impossible because emotion and reason appear to be always intermingled in human cognition. The challenge is in finding the *right balance*. In a sense, the ideal would be to control our passions reasonably

and reliably in relation to the goal of formulating accurate and trustworthy ethical and epistemic judgments and beliefs, instead of letting our passions control our judgments and beliefs.

**129.** For a review of various models of the morally judging mind – including models that have been suggested for how the two systems might interact on the domain of moral judgment – see Guglielmo 2015. Evans & Stanovich (2013, Fig. 1) have also formulated a tripartite model of the (general) dual-processing mind (see also Stanovich 2018a, Fig. 1–4). Furthermore, Pennycook, Fugelsang, & Koehler (2015b, Fig. 1) have presented a three-stage dual-process model of analytic engagement. (see also De Neys 2018; Kahneman 2011.)

**130.** Specifically, this is an area of research in the emerging study of the monitoring and control of reasoning and problem solving; i.e., the study of *meta-reasoning* (see Ackerman & Thompson 2017).

**131.** It is important to note that even though Type 2 processing is much more likely to bring about trustworthy answers in tasks that require the type of non-intuitive, analytic thinking it represents – e.g., possibly in discerning false news headlines, regardless of partisanship (Pennycook & Rand 2019b, 2019c) – it is still of course fallible and can produce different answers for different individuals in different circumstances, for example when guided by biases. Type 2 processing is what is also utilized for *rationalizing* Type 1 sentiments, regardless of their soundness. It merely appears that Type 2 processing is *the best we can do* when confronting complex analytic tasks. The real tragedy, however, is that we often do not realize we are encountering those tasks when most of the time we are engaged in Type 1 autodrives. (Kahneman 2011, 415–418.)

**132.** As far as I’m aware, these terms have not been used in this way before. Of course, they would be subcategories of metacognition (see sect. 6.1.1.1n107).

## Section 6.2

**133.** I think it is reasonable to think of these kinds of considerations as acts (of a sort), given that we are to some significant degree free – in some compatibilist sense – to choose whether to engage in a given (act of) consideration, like the one we are currently engaged in. In any case, the crux of the matter is that relevant considerations during an act (via the engaged, i.e. deliberative perspective) and considerations after the act (via the detached perspective) involve a similar cognitive property (which I happen to be prone to associate with “action”).

**134.** See David B. Wong, *Natural Moralities: A Defense of Pluralistic Relativism* (Oxford University Press, 2009), 234–237, for how ancient Chinese philosopher Zhuangzi may be interpreted to also combine the detached and engaged perspectives in a complimentary manner (though I am not sure if I would accept the overall case for pluralistic relativism). For example, Wong describes Zhuangzi to argue skeptically that detached perspectives go astray as they claim an exclusive and infallible insight into value (236).

Complementary thoughts may also be found in some branches of Buddhism, for example when considering that we may be too attached to the past if we do not realize that our thoughts are actually in the present. To be compassionate, then, it seems to help to be aware of everyone’s engaged perspective often suffering via it being interlinked with their detached perspective, and vice versa. Arguably, this realization, among others, is needed for Buddhist *eudaimonia*. (see, e.g., Dalai Lama 2011, 64–66 & 68–71 & 113–115 & 137–143; Flanagan 2011, 27–29 & 94–98 & 124–126 & 158–159 & 201–202; see also sect. 6.6.2.)

**135.** This may be possible in special cases, though. For example, if short-term memory is malfunctioning, an agent could, theoretically, occupy a perpetual detached perspective of some sort. Or, conversely, if long-term memory is malfunctioning, an agent could, theoretically, occupy a perpetual engaged perspective of some sort. In this context, a regularly functioning memory should be assumed.

**136.** Even though Sher considers the possibility of blaming and holding people responsible being reactions that happen from the perspective of the target agent (see section 3.4.2.1), his consideration does not seem to address the interlinked aspect of the perspectives. Thus, a mistaken presupposition that still seems implicit in Sher’s account – even given that his critique of Nagel’s views would succeed – is that the perspectives are fundamentally distinct (see section 3.4.2.1 & 3.4.2.2).

137. The social psychological bias often called *fundamental attribution error* (abbr. FAE; also known as or closely related to correspondence bias, actor-observer asymmetry, and self-serving bias in attribution) may explain why the interlinked quality of the two perspectives is often glossed over, perhaps especially in individualistic cultures. The phenomenon corresponding with the term “FAE” seems to frustratingly somewhat vary depending on the source, but is here taken to mean the following: The bias refers to people’s potential tendency to attribute their own behavior to external factors, while other people’s behavior is more so used to infer internal dispositions and traits. That is, it refers to the tendency to infer disposition from another actor’s behavior, but not from our own behavior; i.e., to underestimate situational factors and overestimate dispositional factors as something to infer from another actor’s behavior. Insofar as this happens, it may be partly due to the high salience of the outside actor and their striking act as opposed to the low salience of their circumstances. (see Ross 1977 for the first mention of the term FAE; see Bauman & Skitka 2010 for a study on the prevalence among US adult population [53 %] and association with a dispositionist lay philosophy of behavior; see Gawronski 2004, Gilbert & Malone 1995, and Malle 2006 for examples of the discussion about closely related terminology and space of related phenomena with varying amount of support for their existence or prevalence; cf. Dean & Koenig 2019, Masuda & Nisbett 2001 and Kitayama, Ishii, Imada, Takemura, & Ramaswamy 2006a for cultural differences, with collectivistic East Asian cultures displaying partly reduced bias, as compared to individualistic Western cultures; and see Genschow et al. 2017 for an association between overestimating the influence of internal factors of others’ behavior and belief in free will.)

More specifically, however, some important exceptions to the general description of FAE seem to tend to occur when either (1) a perceived enemy or rival does something good, and when (2) a perceived close friend or ally does something bad. In both scenarios, we seem tend to infer external factors while neglecting to consider the role of disposition (Hewstone 1990; Rosenberg & Wolfsfeld 1977). Conversely, when either (3) a perceived enemy or rival does something bad, and when (4) a perceived close friend or ally does something good, we seem tend to infer disposition while neglecting to consider the role of external factors (Hewstone 1990; Rosenberg & Wolfsfeld 1977). This group-level extension or qualification of FAE is sometimes referred to as the “*ultimate attribution error*” or, more fittingly, *intergroup attributional bias* in intergroup causal attribution (Hewstone 1990). It demonstrates our potential in-group-serving tendencies: in a sense, we may be prone to (over)emphasize our group-loyalties even in our intuitive attributions of others’ behavior when group affiliation is salient and the act is perceived to be non-neutral. A meta-analysis has found support for this phenomenon (Hewstone 1990). However, even though the evidence concerning individualistic cultures is comparatively clear, evidence of whether, or how much, there is presence or absence of this group-serving bias in collectivistic cultures appears to be currently mixed, and hence inconclusive (Dean & Koenig 2019).

A further related phenomenon is *group attribution error* (abbr. GAE), referring to two similar phenomena: (1) people’s potential tendency to assume a correspondence between characteristics of an individual group member and the group as a whole, while neglecting variation within the group; and (2) people’s potential tendency to assume a correspondence between a group decision and group members’ attitudes, while neglecting the decision process where many group members can disagree with the result (Allison & Messick 1985; Hamill, Wilson, & Nisbett 1980). In particular, people seem tend to generalize group decisions to group members’ attitudes in the case of out-groups perceived to be threatening, and – similar to FAE – are more likely to infer out-group members’ attitudes from out-group decisions than in-group members’ attitudes from in-group decisions (Allison & Messick 1985; Corneille, Yzerbyt, Rogier, & Buidin 2001). A related phenomenon referred to as the *illusion of out-group homogeneity* (or outgroup homogeneity error) further emphasizes our tendency to falsely perceive out-groups as more homogeneous than they are, as compared to the in-group perceived to be comparatively diverse (Delamater et al. 2015, 458). Collectivistic cultures appear to be even more prone to stereotype people based on group, as compared to individualistic cultures, but may also be more open to change and contradiction in stereotypes, and better recognize groups overlapping (Dean & Koenig 2019). (see also Rosenberg & Wolfsfeld 1977 for how FAE connects with cognitive dissonance theory; sect. 6.1.3.1n125; see also false consensus effect: sect. 6.1.3.1n127.)

Quite strikingly, FAE provides a possible explanation for Sher’s and the possible majority’s intuitions regarding the example cases. Furthermore, it might explain many attributions of moral responsibility manifested on some parts of social media and elsewhere. That is: Sher and the others may be exhibiting FAE when judging the example cases (and others), tending to intuitively associate the negative acts with dispositions (or “constitutive psychology”) of the agents and neglecting to consider the agents’ situation that can be epistemically or otherwise deprived. This appears to impede intergroup empathy and compassion.

Further, *suggested* by research on GAE (though taking some interpretive liberations that may give warrant for caution for the following inference): As Sher and the others view the agents as threatening in terms of breaking what they view to be important moral norms, they are predisposed to associate the actions not only with dispositions (as opposed to external circumstances) but also with generalized features of a threatening out-group. For example, in Sher’s example case *Colicky Baby*, Scout may be intuitively but possibly erroneously connected to a threatening out-group of “agents who give alcohol to babies, *maliciously*” (see sect. 3.3.2; cf. sect. 6.1.1.2). Or, as another example, people on the Internet

who disagree with a particular in-group political viewpoint may, by some, be erroneously (or biasedly) categorized to a threatening political out-group that is exclusively and homogeneously perceived to represent that viewpoint. Even a moderate disagreeing opinion may in this way be mistakenly categorized as an opinion necessarily coming from a member of an extreme out-group, when it conflicts with a view that is perceived to be strongly held within an in-group (see sect. 1.1). That is, GAE – together with the illusion of out-group homogeneity – can not only enable people to falsely perceive all members of an out-group to represent the worst views associated with some highly visible members of the group, they can also enable people to falsely perceive a moderate view resembling some views of some out-group members to necessarily imply membership of the out-group.

Importantly, there is some support for perspective taking training and some forms of mindfulness training reducing FAE (Hooper, Erdogan, Keen, Lawton, & McHugh 2015; Hopthrow, Hooper, Mahmood, Meier, & Weger 2017; see also Catapano, Tormala, & Rucker 2019; Wu & Keysar 2007). Also, expanding the boundaries of *us* to include *them* can positively alter people's views of a perceived out-group (Gaertner & Dovidio 2005; see also Delamater et al. 2015, 460–464; I. H. Smith, Aquino, Koleva, & Graham 2014).

**138.** A distinction between “being responsible” and “taking responsibility” has been made in the literature and utilized vernacularly, but in a bit of a different way than I make and do here (cf. Caruso 2018, ch. 1). By “being responsible” and “taking responsibility”, I am in both cases specifically referring to *moral* responsibility.

**139.** Instead of the usual way of framing the discussion between the engaged and detached perspectives, “the second-person standpoint” could possibly also be utilized. For example, philosopher Stephen Darwall has advocated the concept in his book *The Second-Person Standpoint: Morality, Respect, and Accountability* (2006). Darwall defines the concept as “the [intersubjective] perspective you and I take up when we make and acknowledge claims on one another’s conduct and will” (2006, 3), and builds a whole theory of morality around the concept. Far detached from Darwall’s own conceptualizations and specific context, it seems to me that the second-person standpoint would also need to be understood in roughly the terms I have outlined for it to become pragmatically functional, in the context of the epistemic condition. That is, the relationship between the intersubjective you and I -perspectives ought to also concern us paying appropriate attention to each other’s engaged perspectives. This is, of course, assuming the concept makes sense and is a useful addition to this context in the first place. It seems unclear whether that is the case. (see also Darwall 2007; Lavin 2008.)

**140.** The qualifier “at least partly” is used because there are also usually, if not always, situational external factors at play.

### Section 6.3

**141.** The thought experiments are revisited and trimmed versions of the originals that I developed in spring 2015, as a part of an essay for a course called *Climate Change and Society* (lead by then doctoral-student, now graduated Lauri Lahikainen). During that time, I had not heard of the knowledge requirement, neither from Aristotle, Sher, nor anyone else – and they were not part of the curricula. Thus, these thoughts can be thought to represent my ground-level judgments about the agents in the thought experiments in a way I was inclined to think about them before (philosophical) theory. Spring 2015 was also time before the bulk of the moral outrage phenomenon started to gather wider attention. These points are particularly noteworthy as Sher has expressed that responsibility should satisfy our ground-level, preanalytical judgments (Sher 2009, 152). Even though I’m not sure if I agree, it does seem that our ground-level judgments can be seen to disagree about these cases. Note, though, that already at the time I had a history of studying social and information sciences and psychology for several years, which had (and have) likely affected my cognitive processes of generating and evaluating intuitions. Likewise, of course, no one can escape their past – Sher included. Future, however, will become past soon enough, and may be affected.

**142.** To be fair, I have encountered some highly educated individuals, who for some reason or another seem to misguidedly dispute the overwhelming evidence of AGW (see section 1.2). Nevertheless, the thought experiment expressed in sections 6.3 and 6.4 should guide us in how to approach any individual, on any empirical topic, given that their denial of the relevant evidence is not due to stubborn ideology that denies the evidence despite being well familiar with it (i.e., denialism) but due to genuine ignorance of the relevant facts (i.e., ignorance; in relation to, for example, scientific and/or metacognitive knowledge).

**143.** Thanks goes to philosopher Lauri Lahikainen, who, in the spring of 2015, passingly utilized Bob in an illustration, thus introducing the character to me. Consequently, he inspired the related thought experiments.

**144.** To be sure, if Bob would not for some reason satisfy Sher's criteria of the beliefs being caused by Bob's constitutive features, the premises could be adjusted further. For example, we might add that Bob tends to be loyal to a fault towards the information channels he primarily consumes. This would seem to qualify Bob to lack moral insight or imagination (like the agents in Sher's example cases 7–9; see section 3.3.2).

Further adjustments might fit Bob with Sher's other cases. For example, Bob may have previously investigated AGW through legitimate sources, but after comparing them to the sources of the ideological blogs he consumes, he falsely concluded that the latter had better arguments. This would seem to qualify Bob to display poor judgment (like the agents in Sher's example cases 4–6; see 3.3.2).

One final precaution: even though FEC focuses primarily on *actions* (and *omissions*), I focus primarily on Bob's *beliefs* (and *lack of beliefs*) and only indirectly on his actions. Anything I say could be formulated to focus on Bob's actions, but I think his beliefs are more important because his actions hinge on them. If Sher were merely to hold Bob responsible for his actions, he would not reach the most tangible cause of those actions. It would be roughly analogous to merely focusing our attention to a patient's symptoms instead of the cause of those symptoms. Of course, the beliefs are further caused by the agent's constitutive features, and any relevant external factors, but beliefs are so close to the surface that they allow us to isolate them as an extremely important if not the most significant factor in Bob's actions in question when he is performing them and continuing to perform them, free of worry. If Bob's beliefs about AGW were to come to align with the scientific consensus, his actions and/or his attitude towards his actions would change significantly. At the very least, for his actions to reliably change, it would be especially helpful for his (relevant) beliefs to change first.

**145.** This is a view roughly resembling that of Sher's (see 3.3.2; 4.1.3.2), as well as a view I have seen casually being endorsed in other places (see also Fouke 2012, 120). Notice that AGW is especially fitting an example as practically everyone can be seen to contribute to it – whether or not they are aware of it. However, framing the example in the context of *scientific* knowledge (about AGW) serves only the purpose of the example being topical and easy to approach. It could also be formulated in terms of metacognitive knowledge (see sect. 6.1.1.1n107; 6.4.3.1n179), and possibly other forms of knowledge as well.

**146.** This is a view roughly resembling that of my own (see 3.3.2; 6.1.1). It may also be called *the pragmatic view*, for reasons that will come clear as we progress.

**147.** This applies in different ways to both of the two meanings of “accept responsibility” and “holding responsible” that were introduced in section 6.2.2, and this double application will indeed follow implicitly for the rest of the thesis.

That is, in the view I'm advocating, an agent can properly accept *being responsible* for an act only if they have gained relevant knowledge prior to performing the act, or if they have gained relevant knowledge after performing the act yet do not display other-regarding sorrow and regret. And they can properly accept *taking responsibility* only after having gained relevant knowledge.

Consequently, holding an agent responsible for a negative act in the sense of them being responsible should only occur if they have gained relevant knowledge prior to performing the act, or if they have gained relevant knowledge after performing the act yet do not display other-regarding sorrow and regret. And holding an agent responsible to take responsibility should only occur after they have gained relevant knowledge. If an agent has gained relevant knowledge, they can be held responsible to take responsibility, and they can also be held responsible in the sense of being responsible if they were to act against their previously gained knowledge (given that voluntariness condition was also fulfilled).

A corresponding distinction applies to “assigning responsibility”: assigning responsibility to an agent for a negative act should only occur if they have gained relevant knowledge prior to performing the act, or if they have gained relevant knowledge after performing the act yet do not display other-regarding sorrow and regret. And assigning responsibility to an agent to take responsibility should only occur after they have gained relevant knowledge.

**148.** This view can be seen to be partly motivated via the close relationship between the engaged and the detached perspectives, outlined in section 6.2.1. Perhaps the closest consideration to this that Sher discusses – and dismisses – is the Scanlonian view, outlined in section 3.4.2.3, where Sher criticizes a premise that “no demand can influence an agent's deliberation unless it is accessible to them”. The attributionist Scanlonian view holds that conscious recognition of features of one's situation that provide reasons to do something is required for being conscious of the reasons to do something, and thus, if this is not fulfilled, moral demands cannot fulfill their internal aim of influencing the agent. In the end, Sher dismisses the view only because he doesn't find the reasons Scanlon gives for the view strong enough – albeit he doesn't seem to have a problem with the reasons in any further sense. (see sect. 3.4.2.3; cf. Sher 2009, 66–69.)

**149.** Saying that everyone is responsible for the accumulation of knowledge in humanity would also *ideally* put great responsibility on the individuals who have gained knowledge to spread the knowledge. Note that view 2 goes further and says that not everyone should be held responsible because they don't all have the relevant knowledge; the responsibility should *solely* lie on those who have the knowledge.

**150.** The arguments presented could also be applied to epistemic responsibility as well as to moral responsibility.

**151.** Note that this doesn't take a stance on legal responsibility. Bob may still be legally responsible, if he, for example, neglects paying possible carbon tax even though he does not accept the AGW-related justification behind the tax. Presumably Bob would accept the democratic system, though, thus being willing to obey laws he deems badly justified while being free to campaign against them. (see also sect. 6.4.3.2, question 4.)

**152.** Note that I'm not talking about *foundationalism*, an epistemic theory of justification. Instead, I'm referring to the overall epistemological way of thinking Bob has come to (socially) adopt – his epistemological 'foundation' – whatever the nuances may be.

**153.** Hinge commitments can be described as the underlying *arational* commitments that are needed for rational evaluation to occur. They are to a significant degree socially built, and hence can vary between people. Still, belief change in *deep disagreements* where people have different hinges is arguably possible because there is always *some* overlap. Thus, arguably, diverging hinge commitments can in time change via successful indirect appeals to the overlapping hinges via rational considerations. (Pritchard 2018; see also Ranalli 2018.)

**154.** It is also highly likely that due to living in the sort of environment he does, Bob has not reached a very high level of what in psychology is called *epistemological development*. Educational psychologist Elise J. West (2004) has formulated a four-stage synthesis of available models, the stages including (from lower to higher levels): [1] *absolute* knowing (dependence on perceived authority, with no room for ambiguity), [2] *personal* way of knowing (following a realization of disagreements existing among authorities, concern is focused primarily with perceived evidence that support one's beliefs), [3] *rules-based* knowing (recognition of discipline-specific rules for comparing and judging knowledge claims, but lacking cross-discipline evaluative abilities), and [4] *evaluative* knowing (recognizing the need and taking *responsibility* of continually evaluating all available evidence, experience, and priorities, which may at any time in the future require reconsideration). It seems Bob is currently stuck at Stage 2, at most. The percentage of adults, even those graduating from college, who seamlessly function within the higher stages appears to be low (Richardson 2013). Thus, especially when further combined with, for example, prevailingly low level of science literacy and lack of critical thinking skills, many of us are vulnerable to spurious or premature scientific claims (Kozak 2018).

## Section 6.4

**155.** Whether or not one feels compelled to look up further information about some knowledge claim they hear depends on many social and psychological factors, such as whether or not the agent sees the knowledge claim as important, whether or not the agent is epistemically open to change their views (within the community; e.g., political group, they currently identify with), whether or not the agent remembers the knowledge claim when in a later situation where further information could be looked up, whether or not the agent is in sufficient mental health to be bothered to do any of this, and so on. Many of these factors also relate to whether they have the knowledge and epistemic skills needed to actually find and be able to critically evaluate relevant sources of information; e.g., do they have the necessary *information-seeking skills* and topic-specific literacy (e.g., science and media literacy).

**156.** *Implicit* or indirect assignment of (alleged) responsibility is included, for example, in many value-laden, or judgment-laden, descriptive words and expressions. For example, claiming someone is a "*denialist*" not only asserts – or declares – the label "denialist" to be descriptive of that someone, it also often derogatorily implies responsibility to *not* be a "denialist" or to be *against* being a "denialist" (assuming a serious context). I will not discuss implicit responsibility further in this context but focus on responsibility in more general terms instead. However, it should be noted that there are crucial differences between explicit and various degrees of implicit assignments of responsibility, and while all forms may be important to note in our communication, especially the subtler implicit ones are a more difficult case due to their (perceived) frequency in our communication.

**157.** Similar effect can also be found in situations where the bystander themselves is included in a group that is in potential threat, requiring action, but where taking action is nevertheless inhibited by being in a group of passive people (Latané & Darley 1968). Both phenomena – bystanders being inhibited from providing help for a person in distress, or them being inhibited from taking action in a threatening situation concerning also themselves – are commonly included in the same ‘bystander effect’, as the effect is usually framed from the point of view of the bystander: “an individual’s likelihood of helping decreases when passive bystanders are present in a critical situation” (P. Fischer et al. 2011).

**158.** For a meta-analysis of helping behavior in rural and urban environments, see Steblay 1987; see also Levine, Martinez, Brase, & Sorenson 1994; Levine, Reysen, & Ganz 2008. For notable cross-cultural variation, with Latino/a cultures on average displaying comparatively attenuated effect via their tradition of *simpatia*, see Levine, Norenzayan, & Philbrick 2001 (notably, there appears to be no significant average variation between individualistic and collectivistic cultures). *Simpatia* (in Spanish) or *simpatico* (in Portuguese) refers to a cultural script characteristic of Latin American and Hispanic cultures that emphasize prioritizing amiable social qualities – like friendliness, niceness, agreeableness, and good-naturedness, with helping strangers being part of the script – as compared with, for example, emphasizing achievement and productivity (Levine et al. 2001). However, this is not a moral quality, but a social one, and thus not telling of one’s moral character *per se* (Levine 2003).

**159.** More specifically, these kinds of rationalizations may be thought to lurk behind the reluctance to intervene via *several* social psychological processes suggested to explain the effect. The most prominent three include the *diffusion of responsibility*, *evaluation apprehension*, and *pluralistic ignorance* (P. Fischer et al. 2011; Latané & Nida 1981). Further, a more general *cost/benefit appraisal* could be mentioned (Kerber 1984). Respectively, these suggest people in larger groups, depending on the specific circumstances and individual, may be especially hindered by (i) feeling only fraction of the responsibility in a group than they would alone; (ii) the fear of being negatively evaluated by others; (iii) over-relying on the (lack of) reactions of others in their judgment of the ambiguous, uncertain situation; and/or (iv) in some way being concerned that the costs of intervening outweigh those related to not intervening (P. Fischer et al. 2011; Kerber 1984; Latané and Darley 1970; Latané & Nida 1981). As mentioned, the *arousal-cost-reward model* adds (v) the lack of physiological arousal in a situation to the possible explanatory factors behind the lack of intervention (P. Fischer et al. 2011).

Furthermore, in some situations the bystander(s) may also suffer from (vi) felt incompetence to help (e.g., due to lack of medical education), which prevents them from intervening (Cramer, McMaster, Bartell, & Dragna 1988; Pantin & Carver 1982; though, cf. Latané & Nida 1981, 318; Shotland & Heinold 1985). A novel neuropsychological approach has also suggested that (vii) bystanders may freeze in an emergency situation due to a reflexive emotional reaction dependent on the personality of the bystander (Hortensius & de Gelder 2018). Of course, there can be overlap between these various processes (i)–(vii). (for some less prominent suggested explanations for the bystander effect, see P. Fischer et al. 2011, 518.)

**160.** In addition to the mentioned attenuating factors (danger, perpetrators, and physical costs; P. Fischer et al. 2011), according to the *self-categorization theory*, people are less likely to empathize with and assist perceived out-group members than in-group members, which too seems to play a part in the overall phenomenon in various ways (see Latané & Rodin 1969; Levine, Cassidy, Brazier, & Reicher 2002; Levine & Crowther 2008; see also Bloom 2016, 2017). For example, varying the surrounding groups and their sizes around the victim has, depending on their respective group identity relationships, potentially a negative or a positive effect (Levine & Crowther 2008). Particularly, helping behavior seems less likely to occur when the group identities between the bystander(s) and the victim are not a good and salient match in terms of perceived social or situation-specific norms (in some contexts, a specific, salient group identity match or combination may even facilitate helping the more bystanders there are; Levine & Crowther 2008).

**161.** Latané & Darley (1968, 1970) describe five-step psychological process model to helping, where, interestingly, only the third step is taking responsibility. I see the two first steps relating to knowledge acquisition, required before being able to accept responsibility. All of the five steps they describe are: (1) noticing a critical situation; (2) construing the situation as an emergency; (3) developing a feeling that one has personal responsibility to intervene; (4) evaluating one’s own skills and resources while planning an appropriate course of action; and (5) taking action. Each one of these steps can take a varying amount of time, depending on the individual and circumstances, and each step has the potential of failing or being ignored by the bystander, leading to no help provided (see notes 159 and 160 above for factors that can negatively influence these steps; see also Delamater et al. 2015, 359–363; Frantz & Mayer 2009).

When we consider that knowledge acquisition often requires someone else – like the one in distress – to communicate new knowledge to the bystander successfully, either via explicit words or implicit body language, Latané & Darley’s five steps to helping may also be seen to motivate view 2. That is, when encountering someone we suspect might be in distress,

*we need* to acquire knowledge (and only subsequently take responsibility for further acting accordingly), not leave it to others whose epistemic state is thoroughly unfamiliar to us (i.e., we do not know whether others have succeeded in steps 1–3, with further steps 4 and 5 also being uncertain to succeed). Adopting view 2 better ensures this need is met (as compared to view 1).

**162.** Compare this to the *Doctrine of Doing and Allowing* (e.g., Greene 2013, 240–245): the often persistent cognitive conviction that actively *doing* something X entails greater moral responsibility, as compared to *allowing* something X. For example, throwing a napkin to a bin on a street and missing the throw is usually taken to entail more responsibility to pick up the napkin, as compared to picking up a napkin that was already on the street. In the case of bystander situations, the other's distress is not of our doing, so it is in a similar way psychologically easier to dismiss. But a normative philosophy that targets responsibility to oneself more efficiently (view 2) is more likely to overcome the Doctrine of Doing and Allowing in situations where it should be overcome, as it puts focus on the fact that the continuation of other's distress (or a dirty street etc.) is of your doing insofar as you are actively *baselessly* trusting on others' noticing the situation, defining it as an emergency quickly enough, and accepting responsibility.

Of course, this does not preclude that the doctrine would not be worthy of maintaining in many situations, for example in medical ethics where actively bringing about a patient's death is never acceptable but allowing their death is, in some circumstances. It may even be worthy of maintaining in the napkin case. But bystander situations are not these kinds of situations, requiring quick action.

**163.** Notice how this is akin to the bystander effect: the knowledgeable is a bystander, and the unknowledgeable is the one in epistemic distress, who should be helped, even if it might be challenging. In fact, there is support for the bystander effect occurring among the knowledgeable in situations where an unknowledgeable should be helped: there is research of the phenomenon manifesting in certain chat, e-mail, and forum settings online (Barron & Yechiam 2002; Blair, Thompson, & Wuensch 2005; Markey 2000; van Bommel, van Prooijen, Elffers, & Van Lange 2016; Voelpel, Eckhoff, & Förster 2008). However, as far as I know, the larger problem online has not been studied: people may be wrong together in varying numbers on social media (e.g., in Facebook groups, tightly connected Twitter networks, etc.), with (more) knowledgeable people often dismissing helping these groups (see sect. 1.1 & 1.2).

Measures do seem to be often planned *after* an epistemically misguided group grows to assert power over others (e.g., in the case of countering anti-vaccination movements – see Howard & Reiss 2018 – via targeted information campaigns *after* a spike in vaccine-preventable diseases). At that point, the people who were for a long time in epistemic distress, but sadly dismissed by a wide majority of bystanders, can no longer be ignored. This seems to be consistent with the arousal-cost-reward model, with tangible public distress increasing arousal to address the group (P. Fischer et al. 2011). The passive bystanders here would have been [A] any knowledgeable but silent people in the misguided group, and [B] any knowledgeable people outside the group who were aware of the group yet passed by it. Of course, amidst information campaigns, many may yet remain a passive bystander wherever the campaigns do not reach or are insufficient. Insofar as these sorts of phenomena exist, fed via view 1, I would refer to them as the *epistemic bystander effect*.

In addition to the other factors mentioned in notes 159 and 160 above, in the epistemic cases the possible reluctance to intervene may partly be explained by the misguided group ostracizing outsiders (van Bommel et al. 2016), especially if the group is strongly dogmatic (Nguyen 2018a, 2018b; see also sect. 1.1). For example, doctors and other experts can experience dogmatic insults and even threats if they express pro-vaccination attitudes in news media, let alone if they more directly approach the misguided groups (Karlamañgla 2019). In varying amounts, similar problems are being faced by other professionals on their respective fields, perhaps most notably among climate scientists (e.g., Mann 2016). Further challenge is that many anti-scientific groups are closed (e.g., closed groups on Facebook), and do not allow for dialogue to step into their closed group boundaries (Nguyen 2018a, 2018b). Luckily (perhaps), some social media platforms are becoming more proactive themselves, for example Facebook, Instagram, Pinterest, and YouTube having recently taken actions in an effort to curb anti-vaccination content and to point users towards more trustworthy information sources (Caron 2019). Of course, the risk is that these platforms would become proactive paragons of skewed science if it fits their agenda, but so far this does not seem to be the case.

**164.** This possible plain absurdity in our behavior may also in part explain how moral outrage on social media is perpetuated (see sect. 1.1). Often, we seem to be performing public negative evaluations of perceived out-group agents, merely to enhance our perceived moral status within our perceived in-group (Jordan, Hoffman, Bloom, & Rand 2016; Rothschild & Keefer 2017). What's more, often this happens behind the out-group's back. It needs to be emphasized: epistemically, this helps no one! Hence, any perception of our overall moral status enhancing via this sort of behavior ought to be considered undermined: such a perception seems almost as an illusion within the in-group, deflecting its own communicative shortcomings.

**165.** If you are the victim in a crowd, for example having noticed early symptoms of a stroke, what should you do to best facilitate getting help from bystanders? Simply put: single someone out, do not just generally call for help. That is, assuming you are conscious enough: pick a person, look them in the eye, address them directly, and *assign* them to help you in a precise way (e.g., to call an ambulance). (Cialdini 2007, 138–140.)

Singling someone out in this way can make it very difficult for the person to diffuse their responsibility as it is explicitly assigned to them and to no one else. Further, this direct, precise addressing eliminates ambiguity of the situation for the person, and it potentially brings them into your situational in-group, both increasing the chance of them helping. The fear of being negatively evaluated can also reverse for the person, because if they refuse an explicit call for help directed at them, they might think they could be evaluated more negatively than if they accept. Moreover, to single someone out and affirm the seriousness of the situation can also increase their physiological arousal, hence further increasing the chance of helping behavior. And when the one person helps, it gets the ball rolling and others are more likely to follow. (see note 161 above.)

Of course, in the case of epistemic bystander effect, the “victim” often does not realize they are in epistemic distress, which would emphasize the need for the bystanders to become more proactive. This is an important difference between the classic bystander effect in close-by emergency situations and the potential epistemic bystander effect. Further, as revealed by the Dunning–Kruger effect in the next section (6.4.2.2), the epistemic cases are made even more challenging when noting that the more dogmatic an unknowledgeable agent is, the more convinced they can be that they are knowledgeable and hence *not* in epistemic distress.

**166.** “Absolute” (or “objective”) estimates are estimates that do not require people to compare themselves to others (e.g., when asking how many test questions were answered correctly), whereas relative (or comparative or social) estimates are estimates where performance is evaluated relative to peers (e.g., when asking how well one did relative to a group of people). (e.g., Dunning 2011.)

**167.** This type of phenomenon continues to be replicated and reported in many domains, for example recently in relation to anti-vaccine policy attitudes (Motta, Callaghan, & Sylvester 2018), anti-GMO policy attitudes (Fernbach et al. 2019), analytic reasoning performance (Pennycook, Ross, Koehler, & Fugelsang 2017), and political knowledge (Anson 2018; Hall & Raimi 2018; for a potential attenuating effect of mechanistic and causal understanding of issues, see Fernbach et al. 2013). In some topics – at least in relation to autism and anti-vaccine attitudes, and anti-GMO attitudes – a more radical variant of DKE has been found, which could be called “super DKE” (abbr. SDKE): those who know the least (e.g., anti-vaccine and anti-GMO policy advocates) do not merely think they know substantially more than they do, they even think they know *more than experts* (Fernbach et al. 2019; Motta et al. 2018). These are example topics where there can currently be found an actual average reversal of knowledge and confidence: the less you know, the more you think you know (or, the more knowledge you *perceive* to have – albeit false knowledge – the more confident you are)!

However, unlike in the topics of vaccines and GMOs, SDKE has not been found in relation to climate change attitudes, only DKE (though only in relation to general science knowledge, i.e. science literacy, as domain-specific knowledge was not measured in the case of climate change; Fernbach et al. 2019). The researchers (Fernbach et al. 2019) as well as clinical neurologist and scientific skeptic Steven Novella (2019) have suggested this may be due to the latter topic being more related to political disagreements in relation to proposed solutions (see *solution aversion*: Campbell & Kay 2014), and due to there being widespread bipartisan fear driven by scientific misinformation on the former topics (see Foltz 2018; Howard & Reiss 2018). This seems plausible, especially since science understanding has been associated with vaccine and GMO acceptance but not with AGW acceptance, the latter being best predicted by political conservatism (in the US; Rutjens et al. 2018; see also sect. 1.2). Nevertheless, more research is needed to more confidently and precisely say why or how some topics manifest SDKE while others merely manifest the regular, tamer DKE. Of course, on topics where DKE is found on average, some individuals may still exhibit SDKE. These individuals could also be worthy of studying in the future.

**168.** It should yet be noted that all of the DKE studies mentioned here rely on sample populations from the USA. One may wonder to what degree DKE may be a cultural artifact of, for example, possible comparative cultural trait arrogance or lack of humility, rather than a more universal phenomenon. In fact, a comparative study has indicated that Japanese (collectivistic culture) may be much more likely to *underestimate* their competence, as a consequence of being very open to recognize where to improve; as compared to North Americans (individualistic culture), who are not as sensitive in detecting negative self-relevant information nor as likely to engage in process of self-improvement on a domain after failure, but focus more on domains they are already adept in (Heine et al. 2001). Further, Japanese (compared to Americans) have been found to be much more consistently in a state of “objective self-awareness” where people consider themselves in terms of how they are perceived by others (Heine, Takemoto, Moskalkenko, Lasaleta, & Henrich 2008). Though more cross-cultural research is needed, it may be that socialization in more collectivistic cultures (or in the

specific country of Japan) tends to more likely produce individuals who are metacognitively or otherwise better prepared to recognize their own epistemic shortcomings, compared to socialization in more individualistic cultures (or in North America).

As a further critical point, however, the role of metacognitive differences between the skilled and the unskilled in a given domain as an explanation for DKE has been questioned. Other suggested contributing factors include general biases of self-estimation, performance artefacts, and statistical artefacts, or some combination of all of the above (McIntosh, Fowler, Lyu, & Della Sala 2019). The debate about the explanation is sure to continue, even though at the moment it seems to be one that is largely on the background from the center stage of the actual empirical phenomenon. In this context, I am assuming what seems to be the standard interpretation at least in the most relevant domains, emphasizing the role of metacognition.

**169.** DKE may also play a role in explaining why false news on Twitter spreads much farther, faster, deeper, and more broadly than true news (Vosoughi et al. 2018). This may be because for any given person there are much more areas where they are not experts than areas where they are, and this would likely be reflected in most people's sharing behavior, further guided by confirmation bias (our strong susceptibility to believe stories that support our strongly held beliefs and ideologies). Thus, most people are likely sharing a disproportionate amount of information they are not an expert in and are thus more likely mistaken about – and, at least currently, people seem especially susceptible to share outrage-inducing information (see sect. 1.1 & 1.2).

**170.** Curiously, there are also framing effects suggesting that preformed impressions of expertise might influence overclaiming. It has been found that, independent of actual performance, self-evaluation of a performance in a test can vary merely based on what the test is said to measure (e.g., abstract reasoning vs. computer programming ability), depending on one's preconceptions of ability on the framed area (Ehrlinger & Dunning 2003). This seems to be at least partly because self-perception can alter the way a task is experienced: for example, a task may be experienced as taking less time to fill out if the framing caters one's preconceptions about one's ability (Critcher & Dunning 2009). In a much earlier study, self-perceived expertise was also found to be positively correlated with providing answers to overly difficult questions with feelings of certainty but not positively correlated with answering the questions correctly (Bradley, 1981).

I cannot help but ponder that it may be the case on average, at least in some places or for some individuals, that when self-perceived expertise on a domain of knowledge increases, the willingness to admit uncertainty on that domain decreases, *regardless of warrant*. On some domains, this overconfidence may be encouraged by social expectations set to experts on those domains (e.g., on the domains of financial forecasting, medicine, and scientific research; Kahneman 2011, 261–264).

**171.** It may be that some high-level scientific and/or philosophical skeptics are experts on this domain (depending, for instance, on how you define or conceptualize “knowledge” more precisely, and hence what limits it would more likely have). Interestingly, overconfidence seems to be attenuated already by a cue to “consider [known] unknowns”; steering thoughts away from the usual considerations of only what is perceived to be known (Walters, Fernbach, Fox, & Sloman 2016).

**172.** This is reminiscent of what British philosopher Bertrand Russell also emphasized, perhaps not so kindly, long before the discovery by Dunning and Kruger: “One of the painful things about our time is that those who feel certainty are stupid, and those with any imagination and understanding are filled with doubt and indecision” (Russell 1951, 4–5; see also Russell 2008, 203–204).

**173.** An important nuance to note here is that when view 2 is promoted – i.e., the idea that we should focus on communicating knowledge instead of dealing responsibility – it is equally promoted to Bob and Jack (and everyone else). Thus, as Bob *falsely* thinks he has a lot of knowledge, view 2 promotes him to spread it just as well as it promotes Jack to spread his genuine knowledge. Nevertheless, the view 2 layout is still conducive to better social deliberation than view 1 layout, as the two diverging knowledge claims are dealt without the relative upper hand that Bob might get when view 1 is in effect. And insofar as epistemic virtues and critical thinking have been promoted in addition to view 2, in a semi-Habermasian way it would seem plausible that people siding with Jack would in time – and in quicker time than under view 1 – be able to convince Bob, rather than vice versa. Overall, the social layout for an epistemic community seems to be pragmatically better optimized via view 2 than view 1.

**174.** A similar kind of effect can be seen in the classic social psychology experiment of a participant evaluating lines of different heights in a group that provides wrong answers, first conducted by psychologist Solomon Asch in 1951. According to the standard interpretation, an individual often conforms to – or, more specifically, complies with – a group,

even at the expense of his or her own perception of reality. Although not everyone is as susceptible, and many nuances have become clearer since the 1950s, the basic effect has proven to be statistically very easy to replicate (for a meta-analysis, see Bond & Smith 1996). There is also recently published research indicating that Asch's experiment brings about a similar conformity effect when conducted on social media, on a Facebook profile (Villota & Yoo 2018; see also sect. 1.1).

**175.** Other phenomena – in addition to the bystander effect, Dunning–Kruger effect, and equality bias – that may be interpreted to also support view 2, may include, for example, *moral foundations theory* (Graham et al. 2011, 2013; Haidt 2012; see also note 176; sect. 6.6.1n196), *fundamental attribution error* (see sect. 6.2.1n137), *meta-reasoning* (Ackerman & Thompson 2017), and the importance of *societal trust* (Nguyen 2018a, 2018b; sect. 1.1n7). These are topics that are touched upon elsewhere in this thesis, in other contexts. Further, unmentioned related topics might include, for example, *labeling theory* (see Skaggs 2016), *learned helplessness* (e.g., Landry, Gifford, Milfont, Weeks, & Arnocky 2018), and perhaps *attitude polarization* (e.g., Lord, Ross, & Lepper 1979) (see also Kaufman & Kaufman 2018; Maio, Haddock, & Verplanken 2019; sect. 6.1.3.1). Properly exploring these and any further relevant phenomena in terms of how they may (or may not) support view 2 is a project that would take much more time and space than available here, but for those interested, the previous sections should provide an implicit guideline for continued exploration.

**176.** We tend to avoid utilizing values other than those we ourselves possess because it is cognitively demanding: our natural inclination is to frame matters via our own values, and we tend to be very loyal to them, even in framing. Nevertheless, it would seem to be beneficial to try to reframe moral issues. (Feinberg & Willer 2013, 2015; Hornsey & Fielding 2017; Voelkel & Feinberg 2017; Wolsko et al. 2016; see also Albarracín & Shavitt 2018; Kahan et al. 2017a; Täuber & van Zomeren 2013.)

The research behind this is in part applying the descriptive *moral foundations theory* (MFT) in moral psychology. The research behind the theory seems to have quite successfully demonstrated that values between political groups consistently differ in predictable ways; by different groups sharing differently emphasized psychological 'moral foundations' and their targets. That is why "morality binds and blinds": one political group can be essentially bound by their concerns grounded in their moral foundations while, at the same time, being blind to the concerns of other groups that are bound by differently shaped concerns grounded in theirs (Graham et al. 2009; see also Graham et al. 2013; Haidt 2012).

**177.** I have dedicated a webpage – an online addendum – to pondering some guidelines how we might best communicate with each other on social media and otherwise: <<https://trainingtheelephant.wordpress.com/thesis-addenda/>>.

**178.** As briefly mentioned in 6.3.1n144, Bob's case could also be fitted with other Sher's cases, with little alteration. However, if Sher would for some reason deny the cases of Bob and Jack fitting within his clause 2, it would also pose a problem. For how could then anyone who doesn't share Sher's intuitions decide how to position an agent within FEC clause 2?

**179.** Even though I utilized climate change as an extensive example in the cases of Bob and Jack, it is important to note that the same could be applied to anything that can be imagined concerning (meta)knowledge that the agent lacks. Other examples could be formulated around the cases of, for instance, anti-vaccine or anti-GMO attitudes, homeopathy, conspiracy theories, as well as various topical political questions. Even the case of Alessandra can function as an example: having forgotten her dog in a hot van, if she doesn't afterwards possess enough imagination to realize what proper metacognitive adjustments or heuristics might look like, the help should also come from the outside, and the help would likely most effectively reach its target when done with little to no blaming, or at the very least doing so with emphasis on compassion.

**180.** Philosopher Philip Robichaud has also examined the epistemic condition in the context of AGW, and whether agent's like Bob might be considered responsible or not. Specifically, he examined how volitionist/revisionist and quality-of-will accounts view similar agents (i.e., not Sher's account, which is capacitarian; see sect. 5.4.3–5.4.4.). His final words are reminiscent of the conclusion reached here: the agents would often not be considered responsible, and a duty to reduce ignorance could be undermined. (Robichaud 2016; see also van de Poel, Fahlquist, Doom, Zwart, & Royakkers 2012.)

**181.** Social psychologist Jan-Willem van Prooijen argues in *The Moral Punishment Instinct* (2018, New York, NY: Oxford University Press) that people possess a hard-wired tendency to be emotionally and intuitively motivated to punish those who violate the norms of their group, and rarely think about the matter in any further sense (see also social intuitionist model: sect. 4.1.4n81; sect. 6.1.3.1). If this is the case – as indeed it seems to be – it makes the more rationality-oriented proposition presented here that much more challenging to spread. As van Prooijen also suggests, the punishment

instinct is an early adaptation for humans, having provided strategic survival value for our ancestors in terms of upholding within-group cooperation. However, global intergroup cooperation is a *whole* new ballpark. Thus, the success of the suggestion I am presenting here, and that of intergroup cooperation, would appear to rest on the plausibility of more recent adaptations being able to rise to the challenge.

**182.** Of course, proper care is called for here. For example, it may be that perspective taking training is in some forms or cases prone to exacerbate intergroup conflict via downsides of emotional empathy (cf. Bloom 2016, 2017; Catapano et al. 2019). Thus, it ought to be employed only insofar it can be confidently utilized in a way that cultivates intergroup perspective taking, not only in-group perspective taking. Further research might reveal important nuances to look out for.

Also, both “mindfulness” and “meditation” are quite ambiguous umbrella terms for many different kinds of (meta)cognitive activities, varying and often poorly operationalized in research literature, and hence the overall state of evidence is not clear on the individual nor prosocial benefits (cf. Davidson & Dahl 2018; Van Dam et al. 2018; see also Kozak 2018; Flanagan 2011, 217n7). Still, some forms of mindfulness or meditation could be developed in fruitful ways, for example *compassion meditation* (i.e., *karuna* meditation; CM) and *loving-kindness meditation* (i.e., *metta* meditation; LKM) where the meditator is mentally cultivating virtuous ways of thinking, reacting, and acting by, for example, pre-emptively imagining encounters with others (including intuitively unpleasant people) in difficult situations where compassion is often hard to come by, subsided by one’s selfish side (e.g., Flanagan 2011, 107). (see also sect. 6.6.2 & 6.6.2n199 & 6.6.2n205 & 7.2.2n232.)

In the context of the smaller and what seems to be better confined umbrella of compassion interventions, CM and LKM have been explicitly utilized and found to have individual and, it would seem, potentially prosocial benefits (for an overview and synthesis on compassion interventions, see Kirby 2017). Also, other types of compassion intervention seem to have both individual and potential prosocial benefits, though there is variation between intervention types (for a meta-analysis, see Kirby, Tellegen, & Steindl 2017; see also Kirby 2017). There seems to be some considerable overlap between research on mindfulness, meditation, and compassion interventions, but at the moment the last seems to be the best contained and operationalized group of practices. Some of this discrepancy might be explained when considering that compassion interventions could be thought as a comparatively better handled subcategory of the former two.

My guess would be that some combination of the abovementioned cognitive tools would be a beneficial addition to our educational institutions, to enhance our collective metacognitive processes in order to cultivate pragmatic agent evaluation. More research is needed to find out what that combination might more specifically be, and how it would be safely applied.

**183.** Learning benefits could follow for partly the same reasons as they do when philosophy and ethics is utilized in early formal education. The movement P4C (*Philosophy for Children*) has produced convincing support for the learning benefits of utilizing pupil-led philosophical group dialogues in the classroom. The largest study thus far has been a national randomized 48 school sample from England, consisting of year 4 and 5 primary school students, with 1550 pupils in a trial group (22 schools), receiving the philosophy intervention, and 1609 in a control group (26 schools). The yearlong trial had an aim to “help children become more willing and able to question, reason, construct arguments and collaborate with others”, and it can be considered a success. The results showed that the trial group had improved skills in reading, mathematics, and cognitive attainment (measured via cognitive abilities test, CAT), as compared to the control group. These effects were approximately comparable to two months of additional progress in reading and mathematics (with biggest results among financially disadvantaged pupils), and one month of additional progress in cognitive skills, particularly in verbal skills (with less or no benefit for financially disadvantaged pupils). Feedback from teachers and pupils also suggested improved group cohesion (via increased confidence to speak, and improved listening skills and self-esteem). These are promising results for an intervention of just over a year. A follow-up trial is being expanded to about 200 schools (75 trials, 125 controls; 15,000 year 4, 5, and 6 pupils). The results are to be expected in Spring 2021. (EFF, n.d.; Gorard, Siddiqui, & See 2015.)

Overall, it seems that the benefits in learning performance of the pupils in the trial could be attributed to beneficial effects that doing philosophy can have on metacognitive development and group cohesion. I would suspect that similar learning benefits via metacognitive development and enhanced group cohesion could follow when properly utilizing the abovementioned methods of compassion intervention etc. The most effective solution might thus be to combine philosophy with them. That is, dialogues could be combined with cultivation of intuitions supporting compassionate and pragmatic agent evaluation. Indeed, caution for *only* utilizing dialectic philosophy may be called for, insofar as it can exacerbate intergroup-conflict, for example via implicitly teaching that people who think or act differently to the students (or the most vocal students) in the classroom (i.e., who act non-dialectically and/or non-extrovertly) are to be looked down upon or dismissed.

**184.** Of course, we need not fact-check *all* of our varied beliefs. Insofar as some belief bears little societal or moral importance, it also bears little responsibility. And insofar as we cannot be bothered to check the basis of some of our opinions – for example, due to mental exhaustion or too much felt stress – we can instead choose to *free ourselves* from those opinions and *suspend them* (given that this ability has been enabled by us coming to adopt the kind of inner philosophy that allows for it; one possibility outlined in sect. 6.6.2). If the matter turns out to be important later, we may count on hearing about it from people who know more (or think they know more), especially insofar as the pragmatic view has been widely adopted, and we can always return to examine the matter then.

**185.** Regarding Socratic questioning, see *Street Epistemology* (SE) and related YouTube-videos for a promising and compassionate current day variation that may be utilized anywhere, with little practice: <<https://streetepistemology.com/>> (see also Boghossian 2013). Illustrative videos can be found on, for example, Anthony Magnabosco’s eponymous channel (<<https://www.youtube.com/user/magnabosco210>>) and Reid Nicewonder’s channel *Cordial Curiosity* (<[https://www.youtube.com/channel/UCiWKxPMKUBFjN3Ny\\_VxpkYw](https://www.youtube.com/channel/UCiWKxPMKUBFjN3Ny_VxpkYw)>). The aim of SE is to explore strongly held beliefs of interlocutors during short, friendly and usually one-on-one conversations, often with the effect of the interlocutor’s confidence in those beliefs decreasing via them experiencing *aporia*. This is done with no confrontation but only with calm, curious questions and encouragement of contemplation. On the official website, SE is defined as “a conversational tool that helps people reflect on the reliability of the methods used to arrive at their deeply-held beliefs.”

One may also find the flowchart for argument analysis and evaluation by John Cook, Peter Ellerton, and David Kinkead (2018) a valuable reference, as it summarizes where our inferences can go wrong and gives examples of some of the most common fallacies at different steps of our reasoning processes. This could also be helpful for SE. However, Cook et al. (2018) more particularly suggest us to inoculate people against misinformation by *pre-emptively* exposing them to bad arguments that misinformation rests on and pointing the errors in those arguments.

**186.** Regarding the emotional needs a false belief might serve for a person, the following popular article on *Gizmodo* by Daniel Kolitz, from March 2019, with several expert statements regarding why people believe in pseudoscience, is a nice and quick overview for why compassionate caution is called for: <<https://gizmodo.com/why-do-people-believe-in-pseudoscience-1833193811>>. With similar compassion, and also related to moral reframing, Hornsey & Fielding (2017) provide an illustration for why and how persuasion may be best approached by aligning with (rather than competing with) the “attitude roots” that lay behind people’s beliefs. So far, international attitude roots of anti-vaccination attitudes have been examined more specifically, with according suggestions for communication (see Hornsey, Harris, & Fielding 2018b).

**187.** More specific nuances for the most pragmatic communication are not very clear, although I’ve touched upon some suggestions (see sect. 1.1–1.2 & 6.4.2–6.4.3 & 6.6 & 7.3). It is also possible that different methods work better with different individuals or groups and/or in different situational contexts.

For academic literature that might help in finding better ways to communicate, the emerging science of science communication is something to keep exploring (e.g., K. H. Jamieson, Kahan, & Scheufele 2017). In terms of the Internet, it might be a valuable albeit admittedly “radical” further suggestion to try to also communicate with people in ways where voice or even video is shared, as it may better humanize the participants to each other, compared to mere text-based communication (Cao & Lin 2017; Schroeder et al. 2017). It may also be a worthwhile suggestion to consider how critical thinking could be taught in everyday conversations, even though its teaching is usually studied for the purposes of formal education (for a meta-analysis, see Abrami et al. 2015; see also Tomperi 2017). Street Epistemology may be one approximate model for how that might be achieved (see note 185 above).

Also, see my forthcoming online addendum on the theme of how we might mitigate online outrage, in part via encouraging and modeling better communication habits: <<https://trainingtheelephant.wordpress.com/thesis-addenda/>>.

**188.** One may want to make the distinction between “known unknowns” and “unknown unknowns” as is often done in scientific endeavors. For example, we know that the precise nature of dark matter in cosmology is unknown – but it is a known unknown (to those at least passingly familiar with relevant cosmology). Unknown unknowns are simply, well, unknown, and they are what I refer to here (on the subjective level). Many things are unknown unknowns to our epistemic state until someone effectively enlightens us (notice that one thing can be a known unknown to someone at the same time as it is an unknown unknown to someone else). For example, Jack might not even suspect his communication habits are misguided if the pragmatic view was not introduced to him (see sect. 6.4.1). This is an example of how Jack cannot know what is to him an unknown unknown, even though it is to us at least theorized to be known (provisionally). Similarly, Bob’s error is an example of something that to him is an unknown unknown, though he thinks, due to his error, that his conspiracy theory is known.

**189.** Sher's individualism is further emphasized in his essay collection *Me, You, Us* (2017).

**190.** For an evaluator, Zimmerman's thesis is hard to follow in real life: how often could we confidently come to confirm whether a given transgressive belief or act was traceable to a consciously transgressive one? Philosopher Gideon Rosen (2004), who advocates a similar tracing view as Zimmerman, has argued that we ought to practically always suspend judgments of blame because we can never know the target agent's original or derivative responsibility for an act, and hence can never rule out the possibility that the agent acts from non-culpable ignorance.

**191.** The same could analogously apply to Jack, regarding his communication methods. However, alternatively, it could also be that Jack has *never* heard about his communication methods being unproductive, whereas Bob has heard what he considers misguided claims about AGW. Thus, we may even more straightforwardly consider that Jack's actions do not reflect any blameworthy evaluative judgment on his part.

**192.** It is important to keep in mind that I am solely referring to Smith's attributionism here, and possibly Scanlon's. Given that Sher's view is also a variation of attributionism, there are some substantial differences between the varieties. Eshleman (2016) classifies Smith and Scanlon under the more novel umbrella of *answerability* theories. (see sect. 2.2 & 5.2 & 5.4.)

## Section 6.5

**193.** For some summaries of more relevant research on moral psychology, concerning how situational factors may influence our evaluations of others, see Bloom 2013, 131–157; Haidt 2012; see also sect. 4.1.4n81.

**194.** It may be that there is correlation between descriptive accounts and merit-based views, and normative accounts and consequentialist views of moral responsibility (see ch. 2.2).

## Section 6.6

**195.** This may also partly explain the divisions between philosophers: for example, we may be largely arguing what feels right, in specific example cases and contexts, via our intuitions (cf. sect. 5.4.5; 6.1.3.1). It may have worked fine within limited local groups, but its utility is undermined by the arising need for between-group cooperation. (see Greene 2013, 25–27.) Of course, this is not to say there might not be an ultimate philosophically descriptive answer to be found for our moral intuitions and their connection with responsibility, so long as that answer accounts for the varieties of our intuitions and their possible malleability on a global scale.

**196.** The tragedy of commonsense morality is intended as an analogy to the challenge we are now, in fact, facing. A global connection is beginning to emerge from the process of essentially reconnecting long separated tribes via novel network technologies, economic cooperation, intergovernmental treaties, immigration, and global environmental challenges (see Uz 2015). At the same time, globalists and nationalists vehemently disagree how we should react to these processes (Haidt 2016). There are great hopes for a future of cooperation, but at least as much challenges as well. So far, there have been several approaches that appear to have touched upon this rising challenge that I am aware of, in addition to Greene's deep pragmatism. These are not necessarily all incompatible with each other:

Social psychologist Jonathan Haidt has suggested that an empirically informed Durkheimian form of pluralistic rule utilitarianism, that would be well aware of the socially important dimensions of morality, would be best suited for public policy decisions (2012, 316 & 441n68–n71). At the same time, he views that virtue ethics would be best suited for everyday life (2006, 155–179; 2012, 441n68; see also Haidt 2011b; cf. Greene 2013, 334–346). Further, Haidt (2012) has especially advocated the benefits of us being aware of our different intuitive moral foundations, correlating with political identification, and how it can help us be less self-righteous and better understand and approach one another. I see the moral foundations theory, that Haidt initially developed with Craig Joseph and later with Jesse Graham and others, to have been an especially valuable insight into how to think about group differences and disagreements (see Graham et al. 2009, 2011, 2013; Haidt 2012). There is empirical support for the moral foundations being universal, and similarly variable both within and between populations (Haidt 2012). (see also sect. 4.1.4n81, 6.4.2.4n176.)

Furthermore, psychologist Paul Bloom has not only conducted and summarized a fair amount of research on innate infant morality (that is revealed to be already adaptively biased towards familiarity, for example), but also advocated for the possibilities of rational and compassionate thinking in enabling us to transcend our tribal tendencies (2013, 2016; see

also Flanagan 2017, 261–265; sect. 6.1.1.3n112). Neuroscientist Sam Harris has argued that science ought to form a (normative, utilitarian) foundation for determining moral values connected to well-being of conscious creatures (2010; cf. Born 2014a, 2014b). And, finally, Tenzin Gyatso, the 14th Dalai Lama (2012), has emphasized the virtues of secular ethics, compassion, patience, reason, and science in an interconnected pluralistic world that does not need to be devoid of spirituality, properly understood (see also Flanagan 2011, 140; S. Harris 2014).

Overall, amidst these varied suggestions and possibilities, and others that I am surely unaware of, one thing seems clear: any change that our predicament calls for starts with the individual, like you and me, educating the unknowledgeable, while acknowledging our own fallibility, and cultivating the necessary ways of thinking and acting in ourselves and others.

**197.** Kant, Rawls, and Scanlon represent three distinct deontological thinkers who could be connected to the kind of deontological framing presented here. Approximately described: The German philosopher Immanuel Kant (1724–1804) presented the *categorical imperative* as a moral yardstick that our will ought to follow: an action is a perfect moral duty only if it is something that one could will to be universalized. American philosopher John Rawls (1921–2002) proposed that in structuring a society, our decisions should reflect an *original position* formulated behind a hypothetical *veil of ignorance*: if we are *a priori* unaware of our position in society, the society should be organized in a manner where we would be optimally willing to accept any position within the society. And, the already mentioned American philosopher Thomas Scanlon has proposed via his *contractualism* that actions should be decided on principles that no one would reasonably reject.

**198.** Here, I will extrapolate Flanagan’s definition of eudaimonia<sup>Buddha</sup>, to give a bit fuller view of how he arrives to it:

The foundational *Four Noble Truths* of Buddhism lay behind Flanagan’s definition. These are, charitably interpreted: [1] unsatisfactoriness/pain/suffering (*dukkha*) is abundant in our life and in the lives of others around the world. [2] the cause of *dukkha* is at least in part found in our *first nature*, in the afflictions or defilements (*kleshas*) of our mind. These afflictions are rooted in three poisons: our propensity for (a) thirst or avarice (*lobha* or *raga*), as when we think we need to possess everything we want and that this will make us happy; (b) covetousness, anger, and resentment (*dosa*) at others for what they have and we lack or at others for what they have done that we think affects our happiness; and (c) false beliefs or delusions (*moha*), as when we think things are permanent (see also note 204 below). Commonly, the poison (a) is emphasized, also often described in terms of our propensity for craving or desire for things or temporary states, or clinging, attachment, grasping to or identifying with things/thoughts/views or temporary states. The poison (b) is often also translated as “aversion”. [3] the relief or cessation of suffering is possible if we work to tame or eradicate the afflictions of our mind, cultivating our *second nature*. [4] by following, cultivating, and eventually embodying the *Noble Eightfold Path* we can achieve this relief. The path further consists of cultivating the right view, intention, speech, action, livelihood, effort, mindfulness, and concentration. Of these, the first two pertain to *enlightenment/wisdom*, next three pertain to *virtue/goodness*, and the last three pertain to *meditation/mindfulness* (bodhi (enlightenment) is also often translated, more accurately, as *awakening*). One ought to be careful to not be misled by the bullet point form of the Eightfold Path: what it represents is a roadmap for a long and winding road or path (*dharma*) that is to be followed if eudaimonia<sup>Buddha</sup> is ever to be attained. (Flanagan 2011, 16–22 & 27–29 & 98–107; Wright 2014; for clear video expositions of the Four Noble Truths, and the Noble Eightfold Path, see Kumar 2013a, 2013b.)

Often, it is thought that the goal of the Eightfold Path is *nirvana*: a state of constant highest happiness and quietude, and/or liberation from the *cycle of rebirth (samsara)*. Flanagan intentionally does not much discuss neither nirvana nor rebirth: they are commonly understood to be notoriously unscientific notions, and the most complex and controversial topics across different traditions, and they seem unnecessary for naturalized Buddhism. Arguably, what is often understood by the notion of rebirth is even incompatible with anatman. Similar unscientific confusion surrounds, for example, the concept of *karma* (if understood as, for example, a dependable system of external rewards-and-punishments, further often involving the payoffs being earned in a cycle of conscious rebirth). Flanagan emphasizes that high epistemic standards are important lest we risk fooling ourselves into comfortable delusions (cf. *moha*) that can hinder our way towards intergroup cooperation in the world as it (likely) actually is. I would describe that for us to not go astray in our ethical endeavors, it helps to not accept something without proper epistemic warrant. For example, we would appear to be not well calibrated to recognize all the injustice and unfairness (i.e., parts of *dukkha*) in the world if some part of them could mistakenly be attributed to past lives or deeds of those who in actuality are simply misfortunate. Thus, for naturalized Buddhism, it suffices to simply point out that the Four Noble Truths provide the diagnosis – *dukkha*, caused (at least in part) by the three poisons of our mind – along with the (naturalized) Noble Eightfold Path as the prescription. (Flanagan 2011, 22–23 & 29–35 & 68–80 & 222–225n24 & 226n1–2.)

Still, if one wants to retain these concepts, Flanagan describes how they might be understood in naturalized terms, concerning only *this life* that we do have. Naturalized karmic causation (that Flanagan calls karma<sup>tame</sup>, distinguished from karma<sup>untame</sup>) could be understood in terms of sentient beings simply having abundant effects on other sentient beings and

the environment. We would do well to be better aware of this so we could think, act, and speak more carefully and virtuously, and thus better support eudaimonia<sup>Buddha</sup>. Naturalized nirvana could be understood as the ideal goal of alleviating dukkha by taming the poisons to the maximum degree possible, even if final cessation might be impossible. This results into putting emphasis on (1) enlightenment/wisdom and (2) virtue/goodness and (3) meditation/mindfulness as the ultimate end in themselves (co-constituted, yet analytically separable), providing release from attachment (or at least all unwholesome attachments, as wisdom and virtue may yet be seen as wholesome attachments, constitutive of the dharma path). And, naturalized rebirth could then be understood as having achieved a new beginning by having succeeded at achieving (1) enlightenment (understanding impermanence, dependent origination, anatman, and possibly emptiness), (2) being compassionate, and (3) being mindful. (Flanagan 2011, 22–23 & 72–73 & 131–134 & 211n5; see also Flanagan 2011, 212n6 & 23–35.)

To extrapolate the terminology in Flanagan’s definition of eudaimonia<sup>Buddha</sup>, ‘impermanence’ (*anicca*) refers to a similar thought as that expressed in the aphorisms associated with the ancient Greek philosopher Heraclitus (c. 535 – c. 475 BCE): in Greek “*panta rhei*”, “everything flows”; or “you could not step twice into the same river”. The idea is that all things – material and mental – are in constant flux, always becoming, always transitioning and changing, coming into being and dissolving; and in this sense lack permanence. (Flanagan 2011, 27–29 & 95–96 & 136; Flanagan 2017, 235–237.)

‘Causal interconnectedness’ refers more precisely to ‘dependent origination’ (or ‘dependent being’ or ‘dependent arising’; *pratityasamutpada*) of birth, feeling states, experiences, appearances, world events, death, and all phenomena or *dhammas*. That is, all things are related, interdependent; there is no such thing as independent being; everything that happens depends on other things happening; all things are part of the flux, causally connected to other things, and thus in this sense lack intrinsic (nonrelational) being. (Flanagan 2011, 27–29 & 69 & 126 & 134–136; Flanagan 2017, 235–237.)

In the definition of eudaimonia<sup>Buddha</sup>, Flanagan uses the Sanskrit term *anatman*, but also the Pali term *anatta* is often used (the Pali and Sanskrit terms in Buddhist terminology appear to be often used interchangeably and it seems to vary by practitioner or their tradition which term one prefers to use for a given concept). Commonly, “anatman” is translated as “no-self”, “not-self”, or “non-self”. It describes how there is no self, ego, or soul (*atman*) in the sense that there could be found something in the *five aggregates of clinging* (*skandhas*) that would persist through time and be controlled (these aggregates and the properties of permanence and control appear to co-constitute our commonsense conception of having a self). The aggregates being: [1] form (or matter or body; *rupa*), [2] sensations (or basic feelings, received from form; *vedana*), [3] perceptions (of, for example, identifiable sights or sounds; *samjna*), [4] mental activity or formations (e.g., complex emotions, thoughts, desires, inclinations, habits, decisions, volition; *sankhara*), and [5] consciousness (or awareness, particularly awareness of the contents of the other four aggregates; *vijnana*). In a sense, anatman is what is left of the “self” (or not left) after the deep realization that there is nothing permanent in its nature; that is, after experiencing a kind of “*self-transcendence*” (i.e., experiencing that there is no ‘self’ to transcend or to cling to in the five aggregates). What is left is a psychologically connected, continuous, unfolding and deeply impermanent phenomenon of personality in flow with the (rest of the) universe; a psychological unfolding that is part of a greater unfolding, the Mother of all unfoldings. In this sense, what is commonly referred to by “I” or “self” or “person” is not atman, but anatman. This realization, together with impermanence and dependent origination, helps to alleviate dukkha by providing a new perspective: all bad experiences are but fleeting and we would be misguided to think otherwise of good experiences as well, and, what’s more, the “self” that accompanies those experiences is likewise constantly changing. Thus, any craving or attachment we may feel can be overcome by seeing the world more clearly, including our mind. (Flanagan 2011, 68–70 & 95–99 & 126–127 & 130–131 & 134–139; Wright 2014; Wright 2017, 60–63; see also Flanagan 2011, 123–126 & 159–163; Flanagan 2017, 235–242.)

The thesis of ‘emptiness’ (*sunyata*) is a later expansion of the anatman doctrine from persons to all natural things by Nagarjuna (c. 150 – c. 250 CE), the founding figure of the Buddhist branch of Mahayana (Sanskrit: “Great Vehicle”). Thus, it is not explicitly found in Gautama Buddha’s (c. 563/480 – c. 483/400 BCE) teachings as more closely adhered to by the Buddhist branch of Theravada (Pali: “School of the Elders”). More specifically, ‘emptiness’ follows from the Madhyamaka (Sanskrit: “Middle Path”) school of Mahayana via its critique of the Theravadan view of anatman as not being radical or deconstructive enough in its critique of atman. Essentially, the thesis of emptiness appears to add a reductionist note to all natural things: *everything* is lacking an intrinsic, immutable essence because each thing can be divided into its components, and those into their components, and so on, possibly *ad infinitum* (presumably within the decompositional limits of reality). And – (allegedly) unlike in Theravada – this applies as much to the five aggregates where atman could not be found as it does to atman that dissolved into the aggregates, as the aggregates can be divided into further aggregates, and so on. Thus, atman is not only not found in any of the five aggregates but there is further nothing permanent or independent found in the aggregates of the aggregates, nor in their aggregates, and so on. Anatman, thus understood, dissolves atman not only into the five aggregates where it is not found, but ultimately into emptiness as everything lacks intrinsic (nonrelational) *essences*. Further, what follows from this is even more thorough help in letting go of ego and unhealthy cravings, as anatman – along with everything – is even more thoroughly in flux and relational

than understood in Theravada. Thus, overall, everything is in flux, and all things that seem permanent and that seem to possess intrinsic and nonrelational essences possess only relative, nominal, or pragmatic permanence. (Flanagan 2011, 27–29 & 126–131 & 134–139 & 213–214n11 & 225n25 & 227n5 & 232n13; see also Flanagan 2017, 235–242; see also fundamental attribution error for how unintuitive this kind of thinking seems to be for us: sect. 6.2.1n137.)

Overall, it could be described that the realizations of impermanence, dependent origination, anatman, and possibly emptiness all enhance each other, and they can be made in various gradual orders. And when these realizations have been together deeply absorbed, the enlightenment part of eudaimonia<sup>Buddha</sup> would be satisfied. However, it needs to be yet combined with the remaining two parts – virtue and meditation or mindfulness – as they co-constitute eudaimonia<sup>Buddha</sup>. Arguably, the realizations in the enlightenment part cannot even be deeply absorbed and sufficiently maintained without meditation or mindfulness, gradually learning to pay particular attention to the processes of one’s mind (a distinction can be made between theoretical/intellectual understanding and the more transformative meditative/experiential understanding) (see also note 205 below). And the realization of the importance of the virtues appears to follow from those deeply absorbed realizations as they relate to us being deeply interrelated with all of nature, including all humans. In other words, deep absorption of wisdom follows from meditation, and virtue follows from that wisdom, and meditation or mindfulness also more directly support virtue. In properly fulfilling and continuing to fulfill these three criteria, one can embody the Noble Eightfold Path and become a eudaimon<sup>Buddha</sup> (that is, someone living a life of eudaimonia<sup>Buddha</sup>, taming the three poisons of the mind and thus relieving dukkha in oneself and contributing to relieving it in others). (see also Flanagan 2011, 12 & 16–17 & 29–31; Flanagan 2017, 225–242 & 323n30; Knickelbine 2011.)

**199.** The difference between meditation and mindfulness is often overlooked, though they are closely connected and can support one another. Meditation refers to a comparatively demanding formal sitting or walking practice of deep reflection, or an umbrella of such practices. Mindfulness, on the other hand, is a practice of lived attention that, once learned, can be practiced most of the time – for example, when one consciously metacognitively looks over one’s behavior and thoughts as they continue to arise and subside. (Flanagan 2011, 217n7; see also Flanagan 2011, 105–107.)

**200.** Notably, most Buddhists in Asia do not practice meditation, and earlier teachings of the Buddha – as followed in Theravada Buddhism, for example – seem to be less associated with spirits, protector deities, ghosts, evil spirits, and other supernatural elements that may be found more pronouncedly in some later branches of Buddhism. Also, the Buddha himself was more so deified only by later followers. (Flanagan 2011, 19–20 & 106 & 217n7 & 228n10; D. Smith 2017b.)

**201.** In the English world, dukkha is often translated as “suffering”, but a better descriptive translation would appear to be “unsatisfactoriness” (Flanagan 2011, 99; S. Harris 2014, 38; Wright 2014). On the whole, the term embodies a wide array of unpleasant aspects of human experience, and I therefore prefer and recommend to use the original term as is. There appears to be no straightforward translation for it. (see also definitions for dukkha at SuttaCentral Pali-English dictionaries: <<https://suttacentral.net/define/dukkha>>.)

**202.** For orthodox-challenging pragmatic or instrumental interpretations of anatman, instead of a metaphysical one, see Albahari 2002; Batchelor 2017; cf. note 198 above; Flanagan 2017, 237–239.

**203.** This is at least my Western(ish) perception of the approximate historical *modus operandi* of the Buddhist tradition(s), especially contrasted with those of the much more aggressive Abrahamic traditions. This difference seems to be well reflected in the corresponding teachings of the Buddha versus those in the holy texts of the latter.

Partly demonstrating the historical difference, Buddhist texts seem to have begun to be systematically translated into English only in the late 19th century, and meditative practices in the West started to spread only in the mid-20th century (Cantwell & Kawanami 2009, 95–96; S. Harris 2014, 26). This seems to be comparatively late, considering that the revered over 12,000-page *Pali Canon* (i.e., *Tripitaka*) was likely first written in the 1st century CE after about 500 years of oral tradition following Siddhartha Gautama, i.e. the Buddha. In comparison, the earliest translations of the *Bible* in old English were made approximately in the 7th century CE. (the most affordable resources to the Pali Canon in English, though still incomplete in all formats, appear to be web pages <<https://www.accesstosinsight.org/>> and <<https://suttacentral.net/>>.)

**204.** As introduced above in note 198, the afflictions of the mind are rooted in three poisons in human nature that Buddhist psychology posits, sustaining dukkha and thus deterring the attainment of eudaimonia<sup>Buddha</sup>. Flanagan (2017, 162–164) has also summarized them as follows: “(1) greed, thirst, avarice (*lobha*; *raga*) for all things I want (which is a lot); (2) anger and resentment (*dosa*, *dvesa*) when I don’t get what I want; (3) illusion (*moha*) believing such things as that I deserve to get what I want, and that other beings or the impersonal universe warrant my anger when they don’t deliver”. In *Abhidhamma*, part of the Pali Canon, these are further described to give rise to the “Six Main Mental Afflictions”:

attachment or craving, anger (including hostility and hatred), pridefulness, ignorance and delusion, afflictive doubt, and afflictive views. These in turn are further described to be the roots for the “Twenty Derivative Mental Afflictions”: anger, which comes in five types (wrath, resentment, spite, envy/jealousy, cruelty); attachment, which also comes in five types (avarice, inflated self-esteem, excitation, concealment of one’s own vices, dullness); four kinds of ignorance (blind faith, spiritual sloth, forgetfulness, and lack of introspective attentiveness); and six types of affliction caused by ignorance + attachment (pretension, deception, shamelessness, inconsideration of others, unconscientiousness, and distraction). These may be divided even further. (Flanagan 2011, 101–107; see also Kumar 2014.)

This deep decomposition reveals the magnitude of the task at hand, as these are all obstacles to eudaimonia<sup>Buddha</sup>. Many of these afflictions tend to manifest as reactive mental behavior and action in various situations, often even to our own dismay (cf. reactive attitudes, sect. 2.2; cf. also social intuitionist model, sect. 4.1.4n81). What chiefly sustains these poisons is thought to be the illusory belief that I am an ego, a permanent, unchanging self, atman, and the illusion can be broken when thoroughly realizing that one is not atman but anatman (Flanagan 2017, 162–164; see also Flanagan 2011, 93–114). In that realization, facilitated by meditation and mindfulness, the misguidedness of our many attachments and other afflictions appears to become transparent. (see also note 198 above.)

The process to eliminate or minimize the afflictions can be described as one where our first nature (our given human nature), that contains the poisons, is tamed by our second nature (our cultured nature) by growing the seeds of fellow feeling, (non-discriminatory) empathy, and compassion, and more generally the exceptional virtues and eudaimonia<sup>Buddha</sup> that are more weakly planted in our first nature (Flanagan 2011, 107–114). In other words, the process is one where we relieve dukkha via coming to see the world more clearly, with less ego, which is not our default state.

Similar yet distinct thoughts about cultivating or nourishing our higher nature, while avoiding cultivating or nourishing our lower nature – with disagreements concerning which, if either, is the more natural state – can be found not only from Aristotle (384–322 BCE) and other ancient Greek philosophers, but also from ancient Chinese philosophers who tried to fill in the question about human nature that Confucius (551–479 BCE) had left unanswered. These include philosophers such as Gaozi (c. 420–350 BCE), Mencius (c. 372–289 BCE), Xun Kuang (c. 310 – c. 235 BCE), and Yang Xiong (53 BCE – 18 CE). To some degree similar discussion later occurred in modern Western philosophy: for example, in the textbook discussions about the state of nature between philosophers Thomas Hobbes (1588–1679), John Locke (1632–1704), and Jean-Jacques Rousseau (1712–1778). Overall, it seems to me that the most sensible view is that we have the capacity to cultivate the better angels of our nature, so to speak, but also the capacity to be very easily overcome by both internal and external afflictions. The aim is to find the best means to tame those afflictions, and naturalized Buddhism provides one suggestion. Unlike in the modern Western tradition, the focus here is cultivating the internal (i.e., metacognition), not the external (e.g., social contract).

**205.** I do recommend Owen Flanagan’s book *The Bodhisattva’s Brain: Buddhism Naturalized* (2011) to further approach the Buddhist terminology and way of thinking in terms that a naturalist can appreciate, and, moreover, to see how it compares with, for example, Aristotle’s thinking.

For a secular approach to meditation and mindfulness, I have found Sam Harris’ meditation app *Waking Up* especially helpful: <<https://wakingup.com/>>. The popularized umbrella of “mindfulness meditation” in the West is generally a mixed bag of various kinds of meditation practices that are often detached from wisdom and virtue in eudaimonia<sup>Buddha</sup>. However, with a little digging one can find different meditation practices that are better grounded in the Buddhist traditions (see, e.g., Dalai Lama 2012, 155–183; Flanagan 2011, 105–107 & 194–196; Flanagan 2017, 235–240; S. Harris 2014; Wright 2017; see also sect. 6.4.3.1n182 & 7.2.2n232).

**206.** Flanagan convincingly argues that eudaimonia<sup>Buddha</sup> being best conducive to feeling state happiness<sup>Buddha</sup>, or it being superior to other forms of eudaimonia (e.g., eudaimonia<sup>Aristotle</sup>), cannot be supported empirically, at least currently, as there are too many unclearly defined variables and normative uncertainties – and the same goes for any possible competing suggestions. For example, different forms of eudaimonia have different mental states as their goal, and it is unclear how we ought to normatively compare between them, and especially as practitioners are likely to exhibit confirmation bias towards their own form. Still, these can be argued for (cf. Wright 2017, 269–276). Further, there is some inconclusively ambiguous empirical research that has been conducted thus far, and Flanagan’s exposition of the matter provides helpful guidelines for future research. (Flanagan 2011, 109–114; see also Flanagan 2011, 14–20 & 37–58; Davidson & Dahl 2018; Diebels & Leary 2018; Van Dam et al. 2018.)

**207.** For example, the exceptional virtues along with the traditional virtue of right speech in eudaimonia<sup>Buddha</sup> seem particularly well compatible with the pragmatic view (e.g., D. Smith 2017c, 2017d). Furthermore, eudaimonia<sup>Buddha</sup> may bring other benefits as well:

Consider the cases of Alessandra, Bob and Jack. Had Alessandra nourished a mindful state towards Sheba, for example via karuna and metta meditation, she would have been extra sensitive to the state of Sheba already when leaving the car

and also afterwards. Of course, this would not have necessarily eliminated all memory malfunctions but it would have decreased the odds as she would have been likely to also be less attached to the difficult situation at the school that in the example case distracted her. Or, had Jack nourished a more mindful state towards those who he is prone to hold responsible, he may have been more likely to recognize the counterproductive error of his ways already, by, for example, being more mindful in imagining and approaching the epistemic state of Bob. And, had Bob nourished the virtue of *upekkha*, he may have been eager to find out how he may be mistaken, even if his error was deeply rooted in his hinge commitments (cf. sect. 6.3.3).

In a sense, all of these agents could have been humbler, not attached to a goal or an idea or a mode of action so much so that they failed to notice what is really important for reducing suffering. Alas, all the agents in the original cases – like Sher himself – live in a culture that is apparently prone to not pay that much attention to their metacognition nor that of others (e.g., Dean & Koenig 2019; Heine et al. 2001, 2008; Kitayama et al. 2006a; Proust & Fortier 2018; Wu & Keysar 2007; see also Masuda & Nisbett 2001). We cannot blame them, of course, but we can try to compassionately guide them to a better direction. I know that at least for some it is possible, for my background is similar.

**208.** As, for example, Owen Flanagan and to some degree Dalai Lama have advocated (Flanagan 2011, 61–65 & 214n14; Dalai Lama 2012): we do not need to accept everything the Buddhist tradition posits (or has posited), for example karma (as external cosmic reward for virtuous action) and rebirth/reincarnation (as the same consciousness repeating after death), etc., if it does not fit together with our scientific understanding of the universe. After all, Buddhism at its metaphysical core encourages to accept reality as it most likely is, and one affliction to be overcome is ignorance (Flanagan 2011, 98–101). The emerging form of Buddhism under the umbrella *secular Buddhism* can also be seen to advocate this, Flanagan’s comparative exploration being only one example (Flanagan 2011, 214n14; see also, e.g., Batchelor 2017; Dalai Lama 2012; S. Harris 2014; Wright 2017). Essentially, this allows us to both maintain scientific epistemic integrity, while striving to improve our ways of being in the world.

However, some who endorse similar ideas argue further that the moniker “Buddhism” should not be used at all, but rather we should spread any contemplative insights about the human condition found within via thoroughly nonsectarian means. It is argued that (a) contemplative insights do not in themselves belong to any one tradition and (b) attaching them to religiously connoted “Buddhism”, even if qualified with “secular”, might only confuse the uninitiated (e.g., S. Harris 2006; Wright & Batchelor 2018, 26:51–28:00 & 40:00–41:32). While I agree with the first point, I am currently agnostic on the second, as I see a lot of potential for Buddhism to be thought as a form of secular virtue ethics, though at the same time I can certainly appreciate how many skeptics, religious people, and various institutional actors may currently find the moniker aversive. Simply talking about, for example, “teachings of Buddha” might be a better approach (S. Harris 2006), much like we would talk about the teachings of Socrates, for example. At the same time, as the Buddhist author and former monk Stephen Batchelor suggests, we might want to naturalize the word “dharma” that is used in various ways in the teachings of the Buddha, or to speak of “secular dharma” (though there are problems with this suggestion as well, most notably that the Hindu meaning for dharma as ‘(caste) duty’ is quite different from the Buddhist meaning(s), which might also be a cause for confusion; Wright & Batchelor 2018, 26:51–29:09).

**209.** I should emphasize that endorsing the pragmatic view in terms of the epistemic condition, or in terms of searching for an answer to the tragedy of commonsense morality, is *not* necessarily to endorse a pragmatic theory of truth (at least it is not to endorse a non-Peircean pragmatic theory of truth). Quite the contrary: it is to endorse utilizing our growing empirical and meditative understanding of the human mind to better make our way towards individual, societal, and global well-being, whatever the (fallible and limited) evidence might suggest well-being being constituted of or how we might best get there. In other words, it is to endorse seeking the best *praxis* within the limits our reality permits – including the reality of the functions of our minds – while not fooling ourselves of what that reality most likely includes. In all our endeavors towards widespread well-being, it is important to be psychologically and otherwise realistic.

Thus, it can be understood that truth is not that which works in practice (i.e., pragmatic); but truth (e.g., perceived correspondence or coherence) or scientific/empirical evidence reveals the boundaries of what may work, while minimizing the risk of us fooling ourselves about what those boundaries are. For example, insofar as someone can adopt the pragmatic view, without it requiring a *specific* individual biological human body, that is within the boundaries of our social learning abilities – and arguments along with empirical and experiential evidence can make the case for why we ought to cultivate it in our own character instead of some other guidelines for being in the world.

**210.** See the *Nicomachean Ethics* IV.5 (e.g., Aristotle & Reeve 2014), where Aristotle writes about anger being sometimes praiseworthy.

In the Abrahamic religions – also a prominent tradition in the West – the value of anger does seem to be questioned at times, yet it plays key part in claimed actions of God and is sometimes encouraged in humans as well (Potegal & Novaco 2010). Of course, interpretations and scriptural emphases can always vary. Still, one could argue that the overall

tone of the scriptures is manifesting at least a passive aggressive ethical metaphysics (with “hell” and “heaven” and “(original) sin” and the judging “God”), not to mention the many related and unfoundedly dogmatic “supernatural” epistemological claims transmitted since the decline of antiquity. It might be described that whereas the Abrahamic traditions assert and construct external moderators for human behavior that ought to be followed, the (naturalized) Buddhist tradition suggests to have located internal tendencies that would be in our best interests to examine and overcome (and, arguably, the Abrahamic traditions are products of those tendencies).

**211.** The general sentiment in much of Buddhism is well expressed in a section of the Pali Canon, in *Dhammapada* I.5: “Hatred is never appeased by hatred in this world. By non-hatred alone is hatred appeased. This is a law eternal.”

Also see, for example, the works by Tenzin Gyatso, the 14th Dalai Lama (1940–), on Tibetan Buddhism; and Lucius Annaeus Seneca (c. 4 BCE – 65 CE; especially the work *On Anger (De Ira)*), Epictetus (c. 55–135 CE), and Marcus Aurelius (121–180 CE) on Stoicism. Despite their skepticism towards all dogma, including Stoicist dogma on the possibility of attaining perceptions that conform with objects perceived (in Greek: *phantasia kataleptike*), I would also add Pyrrhonists (and Neo-Pyrrhonists) to the list, whom Flanagan does not mention (see Sextus Empiricus 1996, I.25–30 & III.235–237).

Relatedly, it has been theorized that early Pyrrhonism was heavily influenced by early Buddhism, as Pyrrho of Elis (c. 360–270 BCE) – the namesake and father of Pyrrhonism – may have travelled to India with Alexander the Great (336–323 BCE) (Kuzminski 2008). Philologist Christopher Beckwith (2015) has argued that the one remaining brief passage we have of Pyrrho, via his student Timon of Phlius (c. 320 BCE – c. 235 BCE), can be seen as the earliest known bit of Buddhist doctrinal *text* (in East, the Buddhist tradition was transmitted orally up to the 1st century CE). Beckwith argues that the three key Greek terms present in the text – *adiaphora* (‘without a self-identity’), *astathmeta* (‘unstable, unbalanced, not measurable’), and *anepikrita* (‘unjudged, unfixed, undecidable’) – can be understood to be nearly identical in their meaning as the roots for *anatman* (‘without fixed self’), *dukkha* (‘uneasy, painful, unsatisfactory’), and *anicca* (‘impermanent’) in Buddhism, and that the passage thus supports the hypothesis of Pyrrho–Buddhism connection. Unfortunately, some of Beckwith’s interpretations can be a bit far-fetched from a Buddhist scholar point of view (Jones 2015), and unnecessarily so as his hypothesis could remain without trying to stretch it. Of course, any hard evidence is tough to come by without a lost Pyrrho manuscript surfacing, for example, but the hypothesis is intriguing.

Regardless of whether the hypothesis is correct, in all three schools of thought – Buddhism, Pyrrhonism, Stoicism – one can find what seem to be particularly valuable psychological insights, and each strive towards their own nuanced conception of eudaimonia. Consequently, each of those concepts, along with Aristotle’s eudaimonia, seem worthy of studying and contemplating. One may also find some overlap with successful contemporary therapeutic interventions, perhaps most notably with *Cognitive Behavioral Therapy* (CBT; for a review of meta-analyses and a review of theory, respectively, see Hofmann, Asnaani, Vonk, Sawyer, & Fang 2012; Thoma, Pilecki, & McKay 2015), related *Acceptance and Commitment Therapy* (ACT; for a meta-analysis of randomized controlled trials, see A-Tjak et al. 2015) as well as *Mindfulness-Based Therapies* (MBT; for meta-analyses of applications in therapy and in healthy individuals, respectively, see Khoury et al. 2013; Khoury, Sharma, Rush, & Fournier 2015).

**212.** For example, scientific epistemology and method appear to be something that should be respected, while also respecting naturalized Buddhist virtues that overall more so emphasize the moral goal of alleviating *dukkha* (Flanagan 2009, 59–72). In the Western-originated life stance of *secular humanism* there is considerable overlap between both, but important lessons could still be gained from naturalized Buddhism (D. Smith 2017a; cf. Council for Secular Humanism, n.d.).

The Latino/a tradition of *simpatia* might also be something to examine more closely, even though it is not a moral characteristic *per se* (e.g., Levine et al. 2001; Levine 2003; Rodríguez-Arauz et al. 2018; sect. 6.4.2.1n158). The *ubuntu* philosophy (from the Nguni word roughly meaning ‘humanity’) from sub-Saharan Africa also seems potentially promising in some respects, though it is little systematized, referring to a way of thinking where relations between people are particularly emphasized; a philosophical position often described in terms of a person being a person through other people (e.g., Flanagan 2017, 317n10; Gade 2013; Metz & Gaie 2010). Further, *intellectual humility* has explicitly been theorized to be a virtue (between the vices of intellectual arrogance and intellectual diffidence), and could thus be something to consider and cultivate as such, especially as it has been connected to various prosocial benefits (Krumrei-Mancuso et al. 2019; Samuelson et al. 2015; see also sect. 1.2). See also Peterson and Seligman (2004) for a diverse psychological mapping of (potential) virtues, and D. Jamieson (2007) and Sandberg (2011) for a suggestion to cultivate emission minimizing “green virtues” in the age of AGW. There are sure to be many further ideas worth examining that comparatively few are aware of, also including, for example, in indigenous Amerindian and Oceanic traditions and the already mentioned Stoicism, Pyrrhonism, and ancient Chinese traditions (in notes 204 and 211 above).

Although I have considerable doubts about this, it may even be that anger is in some social/environmental contexts good to express, when expressed for the right things, in the right way, to the right target, at the right time, and for the right

length of time (see Aristotle & Reeve 2014, IV.5; Duhigg 2019). However, this is undermined by suspicions that anger cannot be properly moderated, and hence it may be too risky to ever endorse, plus there may always be better ways to handle situations than anger (e.g., Dalai Lama 2012, 118–119 & 121–122; Seneca 2010; Vernezze 2008; cf. McRae 2015). In any case, even if anger could in some sense be safely endorsed, that would be warranted much less frequently than commonly understood, as indicated by the currently polarized atmosphere online (Brady & Crockett 2019; sect. 1.1). Indeed, incivility, which seems closely connected with (Western) anger, has been found to hinder political persuasion online – with uncivilly presented arguments being perceived as less rational regardless of the level of arguments themselves – and exacerbate polarization (Popan et al. 2019).

**213.** For two reviews of related ideas examining theories of computer-mediated communication (CMC) and how CMC could improve intergroup relations, see Johnson, Lee, Cionea, & Massey 2018; White, Harvey, & Abu-Rayya 2015; see also Cao & Lin 2017.

## Section 6.7

**214.** Philosopher Fred Dretske (2000) has argued that any good epistemic theory ought to leave room for the possibility of beings like us getting it all completely wrong, or as Nguyen (2018c) further puts it: “[t]he brain in the vat will arrive at all the wrong beliefs ... even if they follow all the best epistemic procedures”. I think something along these lines is a good rule of thumb for the moral domain as well. The best we can do may be to carefully formulate theories and heuristics that appear to us to make sense, while accounting for the best available empirical evidence, and then see if they appear to make sense to others as well and see how they compare to what is already out there. We may only hope that future developments would happen in a constructive manner, where all participants try to avoid dogmatism and keep a critical open mind the best they can. Of course, us being fallible humans, the best is what it is.

**215.** The metacognitive clause (1b) is also reminiscent of Zimmerman’s origination thesis (sect. 5.1.2 & 5.4.3), which states that culpability follows only if an act can be traced to an earlier act that the agent considered to be overall morally wrong at the time of the act. However, the metacognitive clause seems to differ from the origination thesis at least in the sense that an agent neglecting to utilize (meta)knowledge – e.g., the metacognitively reinforced Alessandra *consciously* neglecting to utilize the metacognitive heuristic she has gained (sect. 6.1.2) – would not necessarily imply that they believe the omission is overall morally wrong, yet they would still fulfill the metacognitive clause 1b. They may merely think, for example, that “I can make an exception this one time, in these special circumstances that will surely allow no harm done”.

An example could be the metacognitively reinforced Alessandra being in a hurry and thinking it is guaranteed that she will not take long *this one time*. Of course, this would arguably imply that Alessandra still lacks some preceding (meta)knowledge that she’ll surely now learn if she is to get distracted again (hence not fulfilling clause 1b), but it seems equally justified to argue that Alessandra does have the relevant (meta)knowledge yet consciously neglects to follow it (thus fulfilling clause 1b). In either case, assuming 1a has been met, she would seem to have enough relevant (meta)knowledge that she has understood for her to be able to rationally accept responsibility and overall appropriately react if she was to fail and later held responsible. It would seem she would agree if it was then said to her that “you already knew that was not a good idea”, contrary to what would have likely happened before Alessandra became metacognitively reinforced. Thus, the interpretation of clause 1b that allows the metacognitively reinforced Alessandra to be responsible seems to suffice.

**216.** Even though not explicitly discussed in this thesis, it is useful to note that clause 1 implies intentionality, while clause 2 does not. Clause 2 does, however, imply apathy or indifference. Both imply a character with motivational vices.

**217.** Considering that the pragmatic view is consequentialist and not merit-based, it may seem quite pointless to compare it to currently prevailing merit-based views (see sect. 2.2 & 5.4). Still, it can help us locate what the most praiseworthy view of the epistemic condition might currently be outside of the pragmatic view (6.5.2).

Based on the pragmatic view, it would appear that, at least currently, the most praiseworthy view might have the following approximate characteristics: it would be [i] prone to think of responsibility via some form of revisionism or skepticism (and/or, perhaps, attributability or answerability, but being quite dismissive of accountability) (2.2 & 6.4.3.3); [ii] the focus would be on forward-looking views, not backward-looking views (2.2 & 5.4 & 5.4n102); [iii] the contents of awareness that are included in “(meta)knowledge” would be either occurrent and *de dicto* (from clause 1 of the pragmatic view), or dispositional in such a way that the agent does not feel other-regarding sorrow and regret after the (potentially past) disposition and it’s wrongness all-things-considered has been made transparent to them (from clause 2;

see 5.4.1); [iv] the contents of awareness would more largely concern *any* content that would have enabled the agent's epistemic state to perform better in the act in question, and that would likely enable them to perform better in the future (in clause 2; see 5.4.1); and further [v] the most praiseworthy view would be one that would pay particular attention to the processes and effects of our communication and their optimization (6.3 & 6.4); and [vi] the voluntariness of the agent, via a voluntariness condition, would also be paid attention to insofar as it is pragmatic, which, granted, I have not much examined in this thesis, but that I think to be the case based on both first-person experiences of being held accountable of something outside my voluntary control in the sense that external or internal factors prevented me from following my experienced voluntary will, and on third-person observations of people being held accountable in a similar fashion where the results seemed to be deeply unpragmatic.

Hence, considering these requirements, it would appear that the most praiseworthy view outside of the pragmatic view might be some combination of *revisionism* and perhaps *quality-of-will* views (5.4). Thus, out of the views examined in this thesis – not counting the pragmatic view – Michael J. Zimmerman and Angela Smith (and their affiliates) may be most praiseworthy, even though neither of their positions fully satisfy these requirements (cf. 5 & 6.4.3). Out of the examined views, the least praiseworthy would be George Sher's view (cf. 6), though only in a consequentialist sense: it may be praiseworthy in accurately *describing* majority intuitions that, in turn, are not praiseworthy in a consequentialist sense. The same might be said of Smith's account (5.2 & 6.4.3.3). Of course, the pragmatic view itself aims to be the most praiseworthy, in a consequentialist sense (2.2).

Further, some skeptical views, that were not examined, may be found to be even more praiseworthy than the views examined here, outside of the pragmatic view (cf. 6.1.2 & 6.6.3). Relatedly, it may be noted that even though I am not well acquainted with skeptical positions concerning *luck* and issues that it raises (see Caruso 2018, sect. 2.4), it would appear that the pragmatic view might ease some of those concerns. This would appear to be so, because the view minimizes the role of luck as focus is shifted from considering and holding people responsible to pragmatically communicating with them. Under the pragmatic view, the epistemically fortunate would focus on helping the epistemically unfortunate, by aiming to help their epistemic state, instead of eagerly considering them responsible.

## Chapter 7

**218.** It would seem there is persistent tribalism within the human species, as revealed by, for example, prevailing descriptive moral psychology (e.g., Bloom 2013, 2016; Greene 2013; Haidt 2012; see also sect. 1.1 & 1.2 & 6.1.3 & 6.4.2). However, this does not seem to be neither a universal rule nor a rule that could not be mitigated, at least to some significant degree, in propitious societal and educational contexts (e.g., *ibid.*; sect. 6.3 & 6.6.2; Gaertner & Dovidio 2005; I. H. Smith et al. 2014; see also Stanovich 2018b).

### Section 7.1

**219.** Scenarios (2) and (3) also include the possibility of the act being instinctively or unconsciously performed (for example, slipping out a derogatory word in the heat of the moment, or forgetting a dog in car). This can surprise even the agent themselves, if the instinctive act was, for example, an old vice they have long since thought to have overcome, or just something that seems generally out of character. This would implicate a failure in their metacognitive regulation or impulse control. If the agent is generally aware why the specific instinctive performance is negative, and they ought not to have performed it, they would immediately afterwards feel regret (and their metacognitive regulation would likely improve; given no underlying mental health issues that would block metacognitive improvement). If they don't feel regret, and the act was in fact something morally negative, that would implicate the agent either being indifferent to moral considerations or lacking relevant knowledge.

Furthermore, (2) and (3) include the possibility of failure in intent (for example, intending to express a genuine compliment but that the interlocutor surprisingly seems to take as a genuine insult). Our considerations in these sorts of cases would concern both the evaluation of the interlocutor's possibly overblown reaction (e.g., to the intended compliment) and the other agent's act with the failed intent. That is, they concern the question of should either one of them be considered in some way responsible.

**220.** (1a) and (2a) roughly match with Aristotle's voluntary acts, (1b) and (2b) with mixed acts, and (2c) with involuntary acts due to ignorance. (2a) also includes those non-voluntary acts where an agent would be considered responsible, while (2c) includes cases where an agent still appropriately reacts to their non-voluntary act and would thus not be responsible. (see sect. 2.1.1.)

**221.** That is, in the case of the pragmatic view (abbr. PV; sect. 6.7): (1a) would yield responsibility via PV clause 1. (2a) would yield responsibility via PV clause 2.

(1b) and (2b) would yield responsibility via PV clause 1 and 2, respectively, *if and only if* it is the case that the agent was forced to choose between two or more bad choices yet ranked them incorrectly, thus performing a comparatively worse action than what was available, either knowingly when performing the act (in the case of (1b)) or unknowingly but later coming to know of the correct ranking yet not feeling other-regarding regret (in the case of (2b)). If it is the case in (1b) or (2b) that there was no reasonable way to rank the two or more bad choices correctly, there would be no responsibility because the act was then unambiguously involuntary – i.e., done under compulsion – and hence not fulfilling the voluntariness condition. In some cases, responsibility may further remain indeterminated, if there are multiple equally credible yet conflicting views on whether the choices could be ranked (practically, this would imply no responsibility).

In the case of (2c), there is no responsibility following from PV for the past act. In this case, it seems the agent was acting out of genuine ignorance that they are willing to correct via their newly acquired (meta)knowledge, and take subsequent responsibility for their future acts. We should thus compassionately encourage them, not blame them.

Whether (3) would yield responsibility could only be found out by unravelling it into (2a)–(2c) by communicating the knowledge to the target agent. And, finally, (4) and (5) would disqualify the agent from consideration as no ill act was done on their part.

**222.** The chance could also be thought as 3/10 if (3) is further unraveled into (3a)–(3c), corresponding with (2a)–(2c). Or it could be 4/12 if (1b) and (2b) would be divided into resulting in either responsibility, pardon, or indeterminacy; or 6/16 when yet additionally unravelling (3) into (3a)–(3c), with (3b) resulting in either responsibility, pardon, or indeterminacy (cf. notes 220 & 221 above). Thus, all other things being equal, the success rate would more accurately be somewhere between 25–37.5 %. This is, of course, only accounting for the epistemic component in responsibility, not for control or voluntariness. When further accounting for the control or voluntariness condition, the chances would decrease.

Instead of using Aristotle and the pragmatic view as guidelines in formulating the success rate, we could use the searchlight view, Angela Smith's views, or Zimmerman's views, or whatever other theories we may want. None of the three choices that I have examined (searchlight view, Smith, Zimmerman), would appear to much improve the chances. Given that we would need to inspect the target agent's own evaluations of their act (Smith), trace their history of choices relevant to the act (Zimmerman), or simply find out their state of awareness during the act (searchlight view), there would still seem to be at least 8 (or 10, 12, or 16) scenarios to consider and, out of which, around 2 (or 3, 4, or 6) would be ones yielding responsibility. However, at least Zimmerman's tracing would be extremely hard – if not impossible – to follow in practice, especially on social media (cf. Rosen 2004), and Smith's view might also be challenging to follow (cf. sect. 6.4.3.3) Sher's view is not mentioned here because it would not seem to be even potentially normatively praiseworthy an approach, even though it might be descriptively accurate of some people's intuitions (see ch. 5 & 6).

Further theories of responsibility could, of course, affect the chances: for example, all the way from some kind of radical moral responsibility skepticism or thorough revisionism (0 % chance of success) to some kind of radical moral responsibility subjectivism (100 % chance of success). The suggestion here is, all things considered, quite conservative to both directions, though it is more lenient to the direction of skepticism. Views approaching subjectivism seem generally not much endorsed in philosophy, and one can only imagine how catastrophic the state of our communication would be if everyone were to think they can hold anyone responsible and always be correct. (see also sect. 5.4.)

**223.** The severity shift (i.e., increase of outrage in a group of initially outraged individuals) and leniency shift (i.e., decrease of outrage in a group of individuals with initially low levels of outrage) may be explained by: (1) the (in-)group's skewed 'argument pool', essentially enforcing arguments of individuals to only one direction; (2) individuals managing their reputation and preferred self-image within the (in-)group; and (3) the (in-)group facilitating (over)confidence in the rightness of one's views. Additionally, salience plays a role: the deliberation within one's in-group is most salient to oneself (compared to deliberation within out-groups or between groups). (Sunstein 2018, 6–7; see also sect. 1.1 & 1.2.)

Cass Sunstein (2018), who is an American legal scholar, also provides other insights into how contextual variation can lead to changes in individuals' judgments of an otherwise similar case. For example, serendipity (a kind of luck) may be seen to play a significant role in people's beliefs in information cascades – or reputation or outrage cascades – as people are prone to overestimate the amount and/or quality of informational content within a(n in-group) cascade yet still follow the views within what happens to be their in-group (10–11; 11n6).

## Section 7.2

**224.** Alternatively, one might think that instead of suspending judgment, when the impulse first strikes, the emerging task is better thought as trying to *disconfirm* our impulsive judgment before acting on it. However, I think the goal of suspension is better, even though it may be more ambitious, as that would ideally be the mindful state we would conduct our examination in, minimizing procedural biases.

**225.** This would, of course, demand the evaluator to possess requisite information-seeking skills, often including, for example, respect for original sources rather than second-hand accounts, and critical media literacy. Thus, teaching these kinds skills in school and in broader society would also be paramount.

**226.** To determine what knowledge the agent lacks may be less straightforward than one might think, however, when we note the variety in intuitions concerning Sher's example cases (cf. sect. 3.1.2 vs. 6.1.1 vs. ch. 5). The crucial thing to find out is whether there is *any* (meta)knowledge the agent lacked during the act that would have enabled them to act differently had they not lacked that (meta)knowledge (see also sect. 6.7 or ch. 6 more generally).

**227.** In the case of us trying to rule out scenario (5): If what we think of as relevant knowledge seems unfalsifiable (and un-underminable) – that is, if we cannot even imagine new information that could possibly falsify it, nor can we find such imaginations or attempts of falsification elsewhere by our own means – then we cannot test it ourselves. The conclusion in such a case ought to be that we need to talk to those who think otherwise and ask them to provide criteria for falsification and/or sources where they think our knowledge claims have been falsified or undermined. Otherwise, it may merely be that we are missing something we are not aware of. Furthermore, if *no one* can provide criteria for falsification, i.e. if the purported knowledge seems to be thoroughly unfalsifiable, we ought to be quite skeptical of it in the first place (especially if others can make contradictory knowledge claims that are equally unfalsifiable). At the same time, we ought to be vigilant to avoid *ad hoc* explanations (though, if we fail to do so, they are likely to be pointed out by the target agent or third parties, sooner or later; especially if they too follow the pragmatic heuristic).

**228.** “Successfully” here means that the target agent would have become aware of the knowledge and understood it, thus enabling epistemic acceptance of it (see sect. 6.3.1).

**229.** In the case of genuine value disagreements – e.g., when the target agent genuinely views their act to have been overall positive, while the evaluator genuinely views it to have been overall negative – further moral arguments may give warrant to change the evaluation. As thought in the pragmatic view (see PV clause 1a in sect. 6.7), this would constitute as further knowledge – i.e., (meta)knowledge about moral arguments, or “moral knowledge” (cf. Sliwa 2017) – that could, after all, resolve the dispute, and should be communicated accordingly. Specifically, those who view the evaluator's judgment of moral responsibility as misguided should view the evaluator lacking virtue or performing a morally negative act via their evaluation, thus requiring appropriate communication via the pragmatic view and heuristic.

However, insofar as the value disagreements can rest on conflicting and *unhinging* moral foundations between two parties, it may be possible that there are cases of unresolvable disagreement or *deep disagreement* (cf. sect. 6.3.3n153). Nevertheless, the possibility of such a disagreement should not be assumed until all (other) epistemically resolvable issues have been thoroughly examined and ruled out – i.e., relevant knowledge having been mutually communicated. After that, any issues that could be solved by moral arguments should be explored, and then, if even that fails, genuine unresolvable value disagreement would seem to be the inevitable conclusion. Fortunately, it remains unclear whether this sort of structured deliberative discussion would ever reach that point. Even in relatively dire circumstances, it appears that progress towards a mutually respecting solution can be made so long as we retain an ounce of intellectual and moral humility (e.g., Caluwaerts & Deschouwer 2014; Luskin, O'Flynn, Fishkin, & Russell 2014; Nguyen 2011; sect. 6.3.3n153).

**230.** Although I am leaning towards a view that knowledge, in general, is inherently valuable, a nuance that should be noted is that there may still be some situations in which some knowledge ought not always or ever be communicated. In other words, there may be situations where ignorance should be allowed of oneself and/or others. For example, in double blind experiments, both the subjects and the researchers are by design (temporarily) ignorant about who is getting the real medicine on trial and who is getting a placebo (though, they are not ignorant that they are ignorant). Or, movie spoilers are usually frowned upon, as is leaking sensitive scientific knowledge (e.g., knowledge that is bioethically questionable), or sensitive corporate, governmental, or military information. In these kinds of cases, ethical arguments would likely require withholding the knowledge – at least temporarily – or disclosing it strictly on a need-to-know basis. (e.g., S. Miller 2017.)

The kind of knowledge that ought *not* be spread is not relevant here, however, as the focus of the epistemic condition is on knowledge that is morally relevant in the sense it ought to be spread. If some knowledge should not be spread, accepting that assertion would itself rely on (meta)knowledge about ethical arguments that ought to be spread among the concerned parties.

Having said that, there may still be further cases of *deliberate ignorance* or *information avoidance* where individuals choose to not seek knowledge in some more general sense that should perhaps be respected (e.g., Golman, Hagmann, & Loewenstein 2017; Hertwig & Engel 2016). For one example, avoiding potentially unpleasant information might be an understandable and sanctionable way to avoid feeding a depression at times of an already depressive circumstances. These sorts of cases may undermine universalizability of any non-moral culpable ignorance. Still, even if deliberate ignorance is to some degree seen as justified, by following the pragmatic view and heuristic it can be curtailed, since the implication is primarily on the knowledgeable to actively and efficiently (and safely) communicate, not on the unknowledgeable to actively seek guidance.

**231.** Think, for example, the character Aaron Stampler in the movie *Primal Fear* (1996), played by Edward Norton.

**232.** If one finds troublesome, discouraging feelings arising during challenging online discussion – for example, anxiety, frustration, anger, resentment, etc. – meditation and mindfulness can bring some relief. For one point of relief, American journalist and author Robert Wright (2017, 134–136) characterizes how a common mindfulness technique is often summarized by the acronym R.A.I.N.: *Recognize* (the feeling), *Accept* (the feeling, rather than try to drive it away), *Investigate* (the feeling and its relationship to your body), and *Nonidentify* (with the feeling, or notice nonattachment or become nonattached to the feeling). In other words: simply observe the feeling arising, like a puzzling event in the universe, as it is; embrace it as something to be examined from multiple angles, like a scientist would want to carefully examine an external phenomenon, trying to deeply understand it; examine it curiously from an internal distance in the sense that you accept its presence (e.g., notice its predictable or unpredictable causal determinants, its solidity or transparentness, its various constituents, its effect in your body, its effect to your other thoughts arising, the effects of your examination to it, its effects to your behavior towards others, etc.); and notice it changing and eventually subsiding. And don't forget to breathe. It may take some practice to learn to intentionally drop or let go of a negative feeling – or at least tame one substantially quicker than before adopting these kinds of metacognitive techniques – but it can be learned with some effort. (see also sect. 6.6.2, 6.6.2n199, 6.6.2n205.)

## Chapter 8

**233.** It would seem my approach in chapter 6 was roughly the following: First, thoroughly destroy the ego by mercilessly undermining our capabilities as rational beings, demonstrating our incessant and self-destructive propensity for biased and otherwise misguided thinking. Second, from the rubble, build towards something less obsessive of our usual self-centered intuitions that is normatively more desirable and virtuous, and better conducive to well-being. Perhaps thankfully, I'm not sure how much this sort of an approach should be encouraged pedagogically.

**234.** Philosopher Alasdair MacIntyre famously thinks this is the case, starting from the modern era Enlightenment project (1981/1984/2007).

## REFERENCES

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275-314. <https://doi.org/10.3102/0034654314551063>
- Abrams, S. (2016, January). Professors moved left since 1990s, rest of country did not. *Heterodox Academy*. Retrieved from <https://heterodoxacademy.org/professors-moved-left-but-country-did-not/>
- Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8), 607-617. <https://doi.org/10.1016/j.tics.2017.05.004>
- Adams, R. M. (1985). Involuntary sins. *The Philosophical Review*, 94(1), 3-31. <https://doi.org/10.2307/2184713>
- Agarwal, A. (2013, June). *Why massive open online courses (still) matter* [Video]. Retrieved from [https://www.ted.com/talks/anant\\_agarwal\\_why\\_massively\\_open\\_online\\_courses\\_still\\_matter](https://www.ted.com/talks/anant_agarwal_why_massively_open_online_courses_still_matter)
- Albahari, M. (2002). Against no-ātman theories of anattā. *An International Journal of the Philosophical Traditions of the East*, 12(1), 5-20. <https://doi.org/10.1080/09552360220142225>
- Albarracín, D., & Shavitt, S. (2018). Attitudes and attitude change. *Annual Review of Psychology*, 69(1), 299-327. <https://doi.org/10.1146/annurev-psych-122216-011911>
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211-236. <https://doi.org/10.1257/jep.31.2.211>
- Allison, S. T., & Messick, D. M. (1985). The group attribution error. *Journal of Experimental Social Psychology*, 21(6), 563-579. [https://doi.org/10.1016/0022-1031\(85\)90025-3](https://doi.org/10.1016/0022-1031(85)90025-3)
- Amichai-Hamburger, Y. (2005). Internet minimal group paradigm. *CyberPsychology & Behavior*, 8(2), 140-142. <https://doi.org/10.1089/cpb.2005.8.140>
- Anderson, C. (2010, July). *How web video powers global innovation* [Video]. Retrieved from [https://www.ted.com/talks/chris\\_anderson\\_how\\_web\\_video\\_powers\\_global\\_innovation](https://www.ted.com/talks/chris_anderson_how_web_video_powers_global_innovation)
- Anson, I. G. (2018). Partisanship, political knowledge, and the Dunning-Kruger effect. *Political Psychology*, 39(5), 95-97. <https://doi.org/10.1111/pops.12490>
- Aristotle. (1934). *Nicomachean ethics* (H. Rackham, Trans.). Retrieved from <http://www.perseus.tufts.edu>
- Aristotle. (1980). *Nicomachean ethics* (W. D. Ross, Trans.). Retrieved from <http://classics.mit.edu>
- Aristotle, ., & Reeve, C. D. C. (Ed.). (2014). *Nicomachean ethics*. Hackett Publishing Company.
- Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press. <https://doi.org/10.1093/0195152042.001.0001>
- Arpaly, N., & Schroeder, T. (1999). Praise, blame, and the whole self. *Philosophical Studies*, 93(2), 161-188. <https://doi.org/10.1023/A:1004222928272>
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 177-190). Pittsburgh, PA: Carnegie Press.
- Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, 26(8), 1295-1303. <https://doi.org/10.1177/0956797615588195>

- A-Tjak, J. G. L., Davis, M. L., Morina, N., Powers, M. B., Smits, J. A. J., & Emmelkamp, P. M. G. (2015). A meta-analysis of the efficacy of acceptance and commitment therapy for clinically relevant mental and physical health problems. *Psychotherapy and Psychosomatics*, 84(1), 30-36. <https://doi.org/10.1159/000365764>
- Avramova, Y. R., & Inbar, Y. (2013). Emotion and moral judgment. *WIREs: Cognitive Science*, 4(2), 169-178. <https://doi.org/10.1002/wcs.1216>
- Ayer, A. J. (1997). Freedom and necessity. In D. Pereboom (Ed.), *Free Will* (pp. 110-118). Indianapolis, IN: Hackett Publishing Company.
- Bahrami, B. (2018). Making the most of individual differences in joint decisions. In J. Proust & M. Fortier (Eds.), *Metacognitive diversity: An interdisciplinary approach* (chapter 4). Oxford University Press. <https://doi.org/10.1093/oso/9780198789710.003.0004>
- Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *PNAS*, 115(37), 9216-9221. <https://doi.org/10.1073/pnas.1804840115>
- Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130-1132. <https://doi.org/10.1126/science.aaa1160>
- Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, 91(4), 612-625. <https://doi.org/10.1037/0022-3514.91.4.612>
- Ballarini, C., & Sloman, S. A. (2017). Reason and the “motivated numeracy effect.” In Proceedings of the 39<sup>th</sup> Annual Meeting of the Cognitive Science Society, (pp. 1580-1585). Retrieved February 18, 2019, from <https://mindmodeling.org/cogsci2017/papers/0309/paper0309.pdf>
- Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favouritism in cooperation: A meta-analysis. *Psychological Bulletin*, 140(6), 1556-1581. <https://doi.org/10.1037/a0037737>
- Baron, J., & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased? *Perspectives on Psychological Science*, 14(2), 292-303. <https://doi.org/10.1177/1745691618788876>
- Barron, G., & Yechiam, E. (2002). Private e-mail requests and the diffusion of responsibility. *Computers in Human Behavior*, 18(5), 507-520. [https://doi.org/10.1016/S0747-5632\(02\)00007-9](https://doi.org/10.1016/S0747-5632(02)00007-9)
- Batchelor, S. (2017). *Secular Buddhism: Imagining the dharma in an uncertain world*. Yale University Press/New Haven & London.
- Bauman, C. W., & Skitka, L. J. (2010). Making attributions for behaviors: The prevalence of correspondence bias in the general population. *Basic and Applied Social Psychology*, 32(3), 269-277. <https://doi.org/10.1080/01973533.2010.495654>
- Beckwith, C. I. (2015). *Greek Buddha: Pyrrho's encounter with Early Buddhism in Central Asia*. Princeton University Press. <https://doi.org/10.23943/princeton/9780691176321.001.0001>
- Björnsson, G. (2017). Explaining (away) the epistemic condition on moral responsibility. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition* (pp. 146-162). Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.003.0008>
- Blackford, R. (2016, May). The shame of public shaming. *The Conversation*. Retrieved from <https://theconversation.com/the-shame-of-public-shaming-57584>
- Blair, C. A., Thompson, L. F., & Wuensch, K. L. (2005). Electronic helping behaviour: The virtual presence of others makes a difference. *Basic and Applied Social Psychology*, 27(2), 171-178. [https://doi.org/10.1207/s15324834basp2702\\_8](https://doi.org/10.1207/s15324834basp2702_8)
- Blanco, F., & Matute, H. (2018). The illusion of causality: A cognitive bias underlying pseudoscience. In A. B. Kaufman & J. C. Kaufman (Eds.), *Pseudoscience: The Conspiracy Against Science* (pp. 45-75). Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/9780262037426.003.0003>
- Bloom, P. (2013). *Just babies: The origins of good and evil*. The Bodley Head.

- Bloom, P. (2016). *Against empathy: The case for rational compassion*. Ecco.
- Bloom, P. (2017). Empathy and its discontents. *Trends in Cognitive Sciences*, 21(1), 24-31. <https://doi.org/10.1016/j.tics.2016.11.004>
- Boghossian, P. (2013). *A manual for creating atheists*. Pitchstone Publishing.
- Boiger, M., Mesquita, B., Uchida, Y., & Feldman Barrett, L. (2013). Condoned or condemned: The situational affordance of anger and shame in the United States and Japan. *Personality & Social Psychology Bulletin*, 39(4), 540-553. <https://doi.org/10.1177/0146167213478201>
- Bok, H. (1998). *Freedom and responsibility*. Princeton: Princeton University Press.
- Bolin, J. L., & Hamilton, L. C. (2018). The news you choose: News media preferences amplify views on climate change. *Environmental Politics*, 27(3), 455-476. <https://doi.org/10.1080/09644016.2018.1423909>
- Bommarito, N. (2017). Virtuous and vicious anger. *Journal of Ethics and Social Philosophy*, 11(3), 1-27. <https://doi.org/10.26556/jesp.v11i3.112>
- Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, 119(1), 111-137. <https://doi.org/10.1037/0033-2909.119.1.111>
- Booth, C. (2012). Bystanding and climate change. *Environmental Values*, 21(4), 397-416. <https://doi.org/10.3197/096327112X13466893627987>
- Borgesius, F. Z., Trilling, D., Moeller, J., Bodó, B., de Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, 5(1). <https://doi.org/10.14763/2016.1.401>
- Born, R. (2014a, June). *Science, philosophy, & reality* [Blog post]. Retrieved from <https://pointofcontroversy.com/2014/06/09/science-philosophy-reality/>
- Born, R. (2014b, June). *The fight for moral truth* [Blog post]. Retrieved from <https://pointofcontroversy.com/2014/06/20/the-fight-for-moral-truth/>
- Bostrom, N., & Ćirković, M. M. (Eds.). (2008). *Global catastrophic risks*. Oxford University Press.
- Boudry, M., & Pigliucci, M. (Eds.). (2017). *Science unlimited? The challenges of scientism*. The University of Chicago Press. <https://doi.org/10.7208/chicago/9780226498287.001.0001>
- Boutyline, A., & Willer, R. (2016). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, 38(3), 551-569. <https://doi.org/10.1111/pops.12337>
- Bradley, J. V. (1981). Overconfidence in ignorant experts. *Bulletin of the Psychonomic Society*, 17(2), 82-84. <https://doi.org/10.3758/BF03333674>
- Brady, W. J., & Crockett, M. J. (2019). How effective is online outrage? *Trends in Cognitive Sciences*, 23(2), 79-80. <https://doi.org/10.1016/j.tics.2018.11.004>
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *PNAS*, 114(28), 7313-7318. <https://doi.org/10.1073/pnas.1618923114>
- Brenan, M., & Saad, L. (2018, March). Global warming concern steady despite some partisan shifts. *Gallup*. Retrieved from <https://news.gallup.com/poll/231530/global-warming-concern-steady-despite-partisan-shifts.aspx>
- Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, 23(1), 17-33. <https://doi.org/10.1093/jcmc/zmx002>
- Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, 8(1), 108-117. <https://doi.org/10.1016/j.jarmac.2018.09.005>

- Bureau of Labor Statistics, U.S. Department of Labor (2016, June). Employment trends in newspaper publishing and other media, 1990–2016. *The Economics Daily*. Retrieved from <https://www.bls.gov/opub/ted/2016/employment-trends-in-newspaper-publishing-and-other-media-1990-2016.htm>
- Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, 11(2), 215-235. <https://doi.org/10.1007/s11409-015-9142-6>
- Caluwaerts, D., & Deschouwer, K. (2014). Building bridges across political divides: Experiments on deliberative democracy in deeply divided Belgium. *European Political Science Review*, 6(3), 427-450. <https://doi.org/10.1017/S1755773913000179>
- Cameron, C. D., Hutcherson, C. A., Ferguson, A. M., Scheffer, J. A., Hadjiandreou, E., & Inzlicht, M. (2019). Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*. Advance online publication. <https://doi.org/10.1037/xge0000595>
- Campos, A. S. (2013). Responsibility and justice in Aristotle's non-voluntary and mixed actions. *Journal of Ancient Philosophy*, 7(2), 100-121. <https://doi.org/10.11606/issn.1981-9471.v7i2p100-121>
- Campbell, T. H., & Kay, A. C. (2014). Solution aversion: On the relation between ideology and motivated disbelief. *Journal of Personality and Social Psychology*, 107(5), 809-824. <https://doi.org/10.1037/a0037963>
- Cantwell, C., & Kawanami, H. (2009). Buddhism. In L. Woodhead, H. Kawanami, & C. Partridge (Eds.), *Religions in the Modern World*, 2<sup>nd</sup> edition (pp. 68-102). Routledge.
- Cao, B., & Lin, W.-Y. (2017). Revisiting the contact hypothesis: Effects of different modes of computer-mediated communication on intergroup relationship. *International Journal of Intercultural Relations*, 58, 23-30. <https://doi.org/10.1016/j.ijintrel.2017.03.003>
- Capstick, S., Whitmarsh, L., Poortinga, W., Pidgeon, N., & Upham, P. (2015). International trends in public perceptions of climate change over the past quarter century. *WIREs Climate Change*, 6(1), 35-61. <https://doi.org/10.1002/wcc.321>
- Caron, C. (2019, March). Facebook announces plan to curb vaccine misinformation. *The New York Times*. Retrived from <https://www.nytimes.com/2019/03/07/technology/facebook-anti-vaccine-misinformation.html>
- Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology, General*, 148(1), 51-64. <https://doi.org/10.1037/xge0000505>
- Caruso, G. (2018). Skepticism about moral responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2018 Edition. Retrieved from <https://plato.stanford.edu/archives/spr2018/entries/skepticism-moral-responsibility/>
- Caruso, G., & Dennett, D. (2018, October). *Just deserts*. Aeon. Retrieved from <https://aeon.co/essays/on-free-will-daniel-dennett-and-gregg-caruso-go-head-to-head>
- Catapano, R., Tormala, Z. L., & Rucker, D. D. (2019). Perspective taking and self-persuasion: Why “putting yourself in their shoes” reduces openness to attitude change. *Psychological Science*, 30(3), 424-435. <https://doi.org/10.1177/0956797618822697>
- Čavojová, V., Šrol, J., & Adamus, M. (2018). My point is valid, yours is not: Myside bias in reasoning about abortion. *Journal of Cognitive Psychology*, 30(7), 656-669. <https://doi.org/10.1080/20445911.2018.1518961>
- Chandler, K. (2019, March). Watchdog group's leader steps down after founder's firing. *The Associated Press*. Retrieved March 23, 2019, from <https://www.apnews.com/a91fbf8a76944329bf3ed790013727e>
- Chandrashekar, S. P., Yeung, S. K., Yau, K. C., Feldman, G., Chan, M. T. L., Ho, C. M. J., Cheung, C. Y., Lui, C. H., Agarwal, T. K., Wong, C. Y. J., Pillai, T., Fung, H. C., Leung, W. N., Li, Y. T., Tse, C., Cheng, B. L., & Chan, H. Y. C. (2019). *Agency and self-other asymmetries in perceived bias and shortcomings: Replications of the bias blind spot and extensions linking to free will beliefs*. Manuscript submitted for publication. <https://doi.org/10.13140/RG.2.2.19878.16961>

- Christakis, N. A. (2016, June). Teaching inclusion in a divided world. *The New York Times*. Retrieved from <https://www.nytimes.com/2016/06/23/education/teaching-inclusion-in-a-divided-world.html>
- Christakis, N. A. (2019). *Blueprint: The evolutionary origins of a good society*. Little, Brown Spark.
- Cialdini, R. B. (2007). *Influence: The psychology of persuasion*. HarperCollins.
- Cikara, M. (2015). Intergroup schadenfreude: Motivating participation in collective violence. *Current Opinion in Behavioral Sciences*, 3, 12-17. <https://doi.org/10.1016/j.cobeha.2014.12.007>
- Clark, C. J., Ditto, P. H., Shariff, A. F., Luguri, J. B., Knobe, J., & Baumeister, B. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology*, 106(4), 501-513. <https://doi.org/10.1037/a0035880>
- Clark, C. J., Winegard, B. M., & Baumeister, R. F. (2019). Forget the folk: Moral responsibility preservation motives and other conditions for compatibilism. *Frontiers in Psychology*, 10:215. <https://doi.org/10.3389/fpsyg.2019.00215>
- Clark, R. D., & Word, L. E. (1972). Why don't bystanders help? Because of ambiguity? *Journal of Personality and Social Psychology*, 24(3), 392-400. <https://doi.org/10.1037/h0033717>
- Clarke, R. (2014). *Omissions: Agency, metaphysics, and responsibility*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199347520.001.0001>
- Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, 85(5), 808-822. <https://doi.org/10.1037/0022-3514.85.5.808>
- Cohn, N., & Quealy, K. (2019, April). The Democratic electorate on Twitter is not the actual Democratic electorate. *The New York Times*. Retrieved from <https://www.nytimes.com/interactive/2019/04/08/upshot/democratic-electorate-twitter-real-life.html>
- Condon, P., Desbordes, G., Miller, W. B., & DeSteno, D. (2013). Meditation increases compassionate responses to suffering. *Psychological Science*, 24(10), 2125-2127. <https://doi.org/10.1177/0956797613485603>
- Cook, J., Ellerton, P., & Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, 13(2), 024018. <https://doi.org/10.1088/1748-9326/aaa49f>
- Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R. L., Verheggen, B., Maibach, E. W., Carlton, J. S., Lewandowsky, S., Strueck, A. G., Green, S. A., Nuccitelli, D., Jacobs, P., Richardson, M., Winkler, B., Painting, R., & Rice, K. (2016). Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, 11(4), e048002. <https://doi.org/10.1088/1748-9326/11/4/048002>
- Corlett, J. A. (2008). Epistemic responsibility. *International Journal of Philosophical Studies*, 16(2), 179-200. <https://doi.org/10.1080/09672550802008625>
- Corneille, O., Yzerbyt, V. Y., Rogier, A., & Buidin, G. (2001). Threat and the group attribution error: When threat elicits judgments of extremity and homogeneity. *Personality and Social Psychology Bulletin*, 27(4), 437-446. <https://doi.org/10.1177/0146167201274005>
- Council for Secular Humanism. (n.d.). What is secular humanism?. Retrieved April 29, 2019, from <https://secularhumanism.org/what-is-secular-humanism/>
- Cramer, R. E., McMaster, M. R., Bartell, P. A., & Dragna, M. (2006). Subject competence and minimization of the bystander effect. *Journal of Applied Social Psychology*, 18(13), 1133-1148. <https://doi.org/10.1111/j.1559-1816.1988.tb01198.x>
- Critcher, R., & Dunning, D. C. (2009). How chronic self-views influence (and mislead) self-assessments of task performance: Self-views shape bottom-up experiences with the task. *Journal of Personality and Social Psychology*, 97(6), 931-945. <https://doi.org/10.1037/a0017452>
- Crockett, M. J., Özdemir, Y., & Fehr, E. (2014a). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology*, 143(6), 2279-2286. <http://dx.doi.org/10.1037/xge0000018>

- Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014b). Harm to others outweighs harm to self in moral decision making. *PNAS*, 111(48), 17320-17325. <https://doi.org/10.1073/pnas.1408988111>
- Crockett, M. J. (2017a). Moral outrage in the digital age. *Nature Human Behaviour*, 1, 769-771. <https://doi.org/10.1038/s41562-017-0213-3>
- Crockett, M. J. (2017b, October 30). *How social media profits from our moral emotions* [Video]. Retrieved from <https://bigthink.com/videos/molly-crockett-outraged-all-the-time-how-social-media-addicts-us-to-anger>
- Cunningham, M. R., & Baumeister, R. F. (2016). How to make nothing out of something: Analyses of the impact of study sampling and statistical interpretation in misleading meta-analytic conclusions. *Frontiers in Psychology*, 7:1639. <https://doi.org/10.3389/fpsyg.2016.01639>
- Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. M. Doris & the Moral Psychology Research Group (Eds.), *The Moral Psychology Handbook* (pp. 47-71). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199582143.003.0003>
- Dalai Lama. (2012). *Beyond religion: Ethics for a whole world*. Rider.
- Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4), 377-383. <https://doi.org/10.1037/h0025589>
- Darley, J. M., Teger, A. I., & Lewis, L. D. (1973). Do groups always inhibit individuals' responses to potential emergencies? *Journal of Personality and Social Psychology*, 26(3), 395-399. <https://doi.org/10.1037/h0034450>
- Darwall, S. (2006). *The second-person standpoint: Morality, respect, and accountability*. Harvard University Press.
- Darwall, S. (2007). Reply to Korsgaard, Wallace, and Watson. *Ethics*, 118(1), 52-69. <https://doi.org/10.1086/522018>
- DataReportal. (2019, January). *Digital 2019: Global digital yearbook* [Online report]. We Are Social, Hootsuite. Retrieved from <https://datareportal.com/reports/digital-2019-global-digital-yearbook>
- Davidson, R. J., & Dahl, C. J. (2018). Outstanding challenges in scientific research on mindfulness and meditation. *Perspectives on Psychological Science*, 13(1), 62-65. <https://doi.org/10.1177/1745691617718358>
- Dean, K. K., & Koenig, A. M. (2019). Cross-cultural differences and similarities in attribution. In K. D. Keith (Ed.), *Cross-Cultural Psychology: Contemporary Themes and Perspectives, Second Edition* (pp. 575-597). Wiley-Blackwell. <https://doi.org/10.1002/9781119519348.ch28>
- Delamater, J. D., Myers, D. J., & Collett, J. L. (2015). *Social psychology*, 8<sup>th</sup> edition. Westview Press. <https://doi.org/10.4324/9780429493096>
- De Neys, W. (Ed.). (2018). *Dual process theory 2.0*. Routledge. <https://doi.org/10.4324/9781315204550>
- Dennett, D. (1984). *Elbow room: Varieties of free will worth wanting*. Cambridge, MA: MIT Press.
- Dennett, D. (2014, January). Reflections on "Free Will" [Book review]. *Naturalism.Org*. Retrieved from <http://www.naturalism.org/resources/book-reviews/reflections-on-free-will>
- Diebels, K. J., & Leary, M. R. (2018). The psychological implications of believing that everything is one. *The Journal of Positive Psychology*, 1-11. <https://doi.org/10.1080/17439760.2018.1484939>
- Dill, B., & Darwall, S. (2014). Moral psychology as accountability. In J. D'Arms, & D. Jacobson (Eds.), *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics* (pp. 40-83). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198717812.003.0003>
- Ding, D., Maibach, E. W., Zhao, X., Roser-Renouf, C., & Leiserowitz, A. (2011). Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change*, 1, 462-466. <https://doi.org/10.1038/nclimate1295>
- Ditto, P. H., Clark, C. J., Liu, B. S., Wojcik, S. P., Chen, E. E., Grady, R. H., Ceiniker, J. B., & Zinger, J. F. (2019a). Partisan bias and its discontents. *Perspectives on Psychological Science*, 14(2), 304-316. <https://doi.org/10.1177/1745691618817753>

- Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., Celniker, J. B., & Zinger, J. F. (2019b). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, 14(2), 273-291. <https://doi.org/10.1177/1745691617746796>
- Ditto, P. H., & Lopez, D. F. (1992). Motivated scepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63(4), 568-584. <https://doi.org/10.1037/0022-3514.63.4.568>
- Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Psychology of Learning and Motivation*, Vol. 50 (pp. 307-338). Burlington: Academic Press. [https://doi.org/10.1016/S0079-7421\(08\)00410-6](https://doi.org/10.1016/S0079-7421(08)00410-6)
- Dretske, F. (2000). Entitlement: Epistemic rights without epistemic duties? *Philosophy and Phenomenological Research*, 60(3), 591-606. <https://doi.org/10.2307/2653817>
- Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, 9, 111-119. <https://doi.org/10.1038/s41558-018-0360-1>
- Druckman, J. N., Levendusky, M. S., & McLain, A. (2017). No need to watch: How the effects of partisan media can spread via interpersonal discussions. *American Journal of Political Science*, 62(1), 99-112. <https://doi.org/10.1111/ajps.12325>
- Druckman, J. N., Peterson, E., & Slothuus, R. (2013). How elite partisan polarization affects public opinion formation. *American Political Science Review*, 107(1), 57-79. <https://doi.org/10.1017/S0003055412000500>
- Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *PNAS*, 114(36), 9587-9592. <https://doi.org/10.1073/pnas.1704882114>
- Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *The Behavioral and Brain Sciences*, 38(e130). <https://doi.org/10.1017/S0140525X14000430>
- Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5), 729-745. <https://doi.org/10.1080/1369118X.2018.1428656>
- Duhigg, C. (2019, January/February). The real roots of American rage. *The Atlantic*. Retrieved from <https://www.theatlantic.com/magazine/archive/2019/01/charles-duhigg-american-anger/576424/>
- Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one’s own ignorance. In J. Olson & M. P. Zanna (Eds.), *Advances in Experimental Social Psychology* (pp. 247-296). New York: Elsevier. <https://doi.org/10.1016/B978-0-12-385522-0.00005-6>
- Dunning, D. (2014, October). We are all confident idiots. *Pacific Standard*. Retrieved from <https://psmag.com/social-justice/confident-idiots-92793>
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3), 69-106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
- EEF (Education Endowment Foundation). (n.d.). Philosophy for Children (re-grant). Retrieved March 20, 2019, from <https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/philosophy-for-children-effectiveness-trial>
- Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, 84(1), 5-17. <https://doi.org/10.1037/0022-3514.84.1.5>
- Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, 105(1), 98-121. <https://doi.org/10.1016/j.obhdp.2007.05.002>

- Eibenberger, S., Gerlich, S., Arndt, M., Mayor, M., & Tüxen, J. (2013). Matter-wave interference of particles selected from a molecular library with masses exceeding 10 000 amu. *Physical Chemistry Chemical Physics*, 15(35), 14696-14700. <https://doi.org/10.1039/c3cp51500a>
- Eitan, O., Viganola, D., Inbar, Y., Dreber, A., Johannesson, M., Pfeiffer, T., Thau, S., & Uhlmann, E. L. (2018). Is research in social psychology politically biased? Systematic empirical tests and a forecasting survey to address the controversy. *Journal of Experimental Social Psychology*, 79, 188-199. <https://doi.org/10.1016/j.jesp.2018.06.004>
- Epstein, Z., Pennycook, G., & Rand, D. (2019, April 9). Letting the crowd steer the algorithm: Laypeople can effectively identify misinformation sources. PsyArXiv Preprint. <https://doi.org/10.31234/osf.io/z3s5k>
- Eshleman, A. (2016). Moral responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2016 Edition. Retrieved from <https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/>
- Eskine, K. J., Kaciniak, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustory disgust influences moral judgment. *Psychological Science*, 22(3), 295-299. <https://doi.org/10.1177/0956797611398497>
- European Commission, Directorate-General for Climate Action. (2017). *Special Eurobarometer 459: Report – Climate change*. Retrieved from [https://ec.europa.eu/clima/sites/clima/files/support/docs/report\\_2017\\_en.pdf](https://ec.europa.eu/clima/sites/clima/files/support/docs/report_2017_en.pdf)
- Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-241. <https://doi.org/10.1177/1745691612460685>
- Fantl, J. (2017). Knowledge how. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2017 Edition. Retrieved from <https://plato.stanford.edu/archives/fall2017/entries/knowledge-how/>
- Faraci, D., & Shoemaker, D. (2014). Huck vs. JoJo: Moral ignorance and the (a)symmetry of praise and blame. In J. Knobe, T. Lombrozo, & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy, Vol. 1* (pp. 7-27). Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198718765.003.0002>
- Farrell, J., McConnell, K., & Brulle, R. (2019). Evidence-based strategies to combat scientific misinformation. *Nature Climate Change*, Perspective. <https://doi.org/10.1038/s41558-018-0368-6>
- Fastame, M. C., & Penna, M. P. (2013). Psychological well-being and metacognition in the fourth age: An explorative study in an Italian oldest old sample. *Aging & Mental Health*, 18(5), 648-652. <https://doi.org/10.1080/13607863.2013.866635>
- Fatke, M. (2017). Personality traits and political ideology: A first global assessment. *Political Psychology*, 38(5), 881-899. <https://doi.org/10.1111/pops.12347>
- Fazio, L., Rand, D., & Pennycook, G. (2019, March 1). Repetition increases perceived truth equally for plausible and implausible statements. PsyArXiv Preprint. <https://doi.org/10.31234/osf.io/qys7d>
- Federico, C. M., & Malka, A. (2018). The contingent, contextual nature of the relationship between needs for security and certainty and political preferences: Evidence and implications. *Political Psychology*, 39(S1), 3-48. <https://doi.org/10.1111/pops.12477>
- Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science*, 24(1), 56-62. <https://doi.org/10.1177/0956797612449177>
- Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin*, 41(12), 1665-1681. <https://doi.org/10.1177/0146167215607842>
- Feltz, A., & Millan, M. (2015). An error theory for compatibilist intuitions. *Philosophical Psychology*, 28(4), 529-555. <https://doi.org/10.1080/09515089.2013.865513>
- Fernbach, P. M., Light, N., Scott, S. E., Inbar, Y., & Rozin, P. (2019). Extreme opponents of genetically modified foods know the least but think they know the most. *Nature Human Behaviour*, 3(3), 251-256. <https://doi.org/10.1038/s41562-018-0520-3>
- Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, 24(6), 939-946. <https://doi.org/10.1177/0956797612464058>

- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.
- Festinger, L. (1962, October). Cognitive dissonance. *Scientific American*. <https://doi.org/10.1038/scientificamerican1062-93>
- Fetterman, A. K., Curtis, S., Carre, J., & Sassenberg, K. (2019). On the willingness to admit wrongness: Validation of a new measure and an exploration of its correlates. *Personality and Individual Differences*, 138(1), 193-202. <https://doi.org/10.1016/j.paid.2018.10.002>
- Fetterman, A. K., Muscanell, N., Covarrubias, J. J., & Sassenberg, K. (2018). *When you are wrong, just admit it: Wrongness admission leads to better interpersonal impressions*. (Manuscript submitted for publication).
- FIRE (Foundation for Individual Rights in Education). (2016). *Yale University: Protesters at Yale threaten free speech, demand apologies and resignations from faculty members over Halloween email* [Case Summary and Materials]. Retrieved from <https://www.thefire.org/cases/protesters-at-yale-threaten-free-speech-demand-apologies-and-resignations-from-faculty-members-over-halloween-email/>
- Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control*. Cambridge: Cambridge University Press.
- Fischer, J. M., & Tognazzini, N. A. (2009). The truth about tracing. *Noûs*, 43(3), 531-556. <https://doi.org/10.1111/j.1468-0068.2009.00717.x>
- Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, 137(4), 517-537. <https://doi.org/10.1037/a0023304>
- FitzPatrick, W. J. (2008). Moral responsibility and normative ignorance: Answering a new skeptical challenge. *Ethics*, 118(4), 589-613. <https://doi.org/10.1086/589532>
- FitzPatrick, W. J. (2017). Unwitting wrongdoing, reasonable expectations, and blameworthiness. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition* (pp. 29-46). Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.003.0001>
- Flanagan, O. (2011). *The bodhisattva's brain: Buddhism naturalized*. MIT Press. <https://doi.org/10.7551/mitpress/7414.001.0001>
- Flanagan, O. (2017). *The geography of morals: Varieties of moral possibility*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780190212155.001.0001>
- Flanagan, O., & Williams, R. A. (2010). What does the modularity of morals have to do with ethics? Four moral sprouts plus or minus a few. *Topics in Cognitive Science*, 2(3), 430-453. <https://doi.org/10.1111/j.1756-8765.2009.01076.x>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906-911. <https://doi.org/10.1037//0003-066x.34.10.906>
- Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1), 298-320. <https://doi.org/10.1093/poq/nfw006>
- Foa, R. S., & Mounk, Y. (2016). The democratic disconnect. *Journal of Democracy*, 27(3), 5-17. <https://doi.org/10.1353/jod.2016.0049>
- Folta, K. M. (2018). Food-o-science pseudoscience: The weapons and tactics in the war on crop biotechnology. In A. B. Kaufman & J. C. Kaufman (Eds.), *Pseudoscience: The Conspiracy Against Science* (pp. 103-135). Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/9780262037426.003.0005>
- Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, 5, 5-15. Retrieved May 13, 2019, from <https://philpapers.org/rec/FOOTPO-2>
- Fouke, D. C. (2012). Blameworthy environmental beliefs. *Environmental Ethics*, 34(2), 115-134. <https://doi.org/10.5840/enviroethics201234211>
- Frankfurt, H. G. (1988). *The importance of what we care about*. Cambridge: Cambridge University Press.

- Frantz, C. M., & Mayer, F. S. (2009). The emergency of climate change: Why are we failing to take action. *Analysis of Social Issues and Public Policy (ASAP)*, 9(1), 205-222. <https://doi.org/10.1111/j.1530-2415.2009.01180.x>
- Friesen, J. P., Campbell, T. H., & Kay, A. C. (2015). The psychological advantage of unfalsifiability: The appeal of untestable religious and political ideologies. *Journal of Personality and Social Psychology*, 108(3), 515-529. <https://doi.org/10.1037/pspp0000018>
- Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, 72, 1-12. <https://doi.org/10.1016/j.jesp.2017.04.003>
- Gade, C. B. N. (2012). What is ubuntu,? Different interpretations among South Africans of African descent. *South African Journal of Philosophy*, 31(3), 484-503. <https://doi.org/10.1080/02580136.2012.10751789>
- Gaertner, S. L., & Dovidio, J. F. (2005). Categorization, recategorization, and intergroup bias. In J. F. Dovidio, P. Glick, & L. A. Rudman (Eds.), *On the Nature of Prejudice: Fifty Years after Allport* (pp. 71-88). Blackwell Publishing Ltd. <https://doi.org/10.1002/9780470773963.ch5>
- Gardner, M. (1957). *Fads & fallacies in the name of science*, 2<sup>nd</sup> edition (reprint). Dover.
- Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018, April). *Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship*. Paper presented at the World Wide Web Conference, Lyon, France. <https://doi.org/10.1145/3178876.3186139>
- Garrett, R. K., & Weeks, B. E. (2017). Epistemic beliefs' role in promoting misperceptions and conspiracist ideation. *PLOS ONE*, 12(9), e0184733. <https://doi.org/10.1371/journal.pone.0184733>
- Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the United States, 1974 to 2010. *American Sociological Review*, 77(2), 167-187. <https://doi.org/10.1177/0003122412438225>
- Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*, 15(1), 183-217. <https://doi.org/10.1080/10463280440000026>
- Genschow, O., Rigoni, D., & Brass, M. (2017). Belief in free will affects causal attributions when judging others' behaviour. *PNAS*, 114(38), 10071-10076. <https://doi.org/10.1073/pnas.1701916114>
- Gershoff, E. T., & Grogan-Kaylor, A. (2016). Spanking and child outcomes: Old controversies and new meta-analyses. *Journal of Family Psychology*, 30(4), 453-469. <https://doi.org/10.1037/fam0000191>
- Gholipour, B. (2017, May). New AI tech can mimic any voice. *Scientific American*. Retrieved from <https://www.scientificamerican.com/article/new-ai-tech-can-mimic-any-voice/>
- Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, 117(1), 21-38. <https://doi.org/10.1037/0033-2909.117.1.21>
- Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: Free Press.
- Ginet, C. (2000). The epistemic requirements for moral responsibility. *Noûs*, 34(s14), 267-277. <https://doi.org/10.1111/0029-4624.34.s14.14>
- Glover, J. (1970). *Responsibility*. London: Routledge and Kegan Paul.
- Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1), 96-135. <https://doi.org/10.1257/jel.20151245>
- Gorard, S., Siddiqui, N., & See, B. H. (2015). Philosophy for Children: SAPERE, evaluation report and executive summary. EEF (Education Endowment Foundation). Retrieved March 19, 2019, from <https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/philosophy-for-children>
- Goreis, A., & Voracek, M. (2019). A systematic review and meta-analysis of psychological research on conspiracy beliefs: Field characteristics, measurement instruments, and associations with personality traits. *Frontiers in Psychology*, 10:205. <https://doi.org/10.3389/fpsyg.2019.00205>

- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55-130. <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>
- Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5), 1029-1046. <https://doi.org/10.1037/a0015141>
- Graham, J., Nosek, B. A., & Haidt, J. (2012). The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PLOS ONE*, 7(12), e50092. <https://doi.org/10.1371/journal.pone.0050092>
- Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2), 366-385. <https://doi.org/10.1037/a0021847>
- Graton, A., Ric, F., & Gonzalez, E. (2016). Reparation or reactance? The influence of guilt on reaction to persuasive communication. *Journal of Experimental Social Psychology*, 62, 40-49. <https://doi.org/10.1016/j.jesp.2015.09.016>
- Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. The Penguin Press.
- Greene, J. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, 124(4), 695-726. <https://doi.org/10.1086/675875>
- Guéguen, N. (2014). Commitment and the responsive bystander: A field evaluation with a less problematic request. *Psychological Reports*, 115(2), 607-611. <https://doi.org/10.2466/17.07.PR0.115c18z9>
- Guéguen, N., Dupré, M., Georget, P., & Sénémeaud, C. (2014). Commitment, crime, and the responsive bystander: Effect of the commitment form and conformism. *Psychology, Crime & Law*, 21(1), 1-8. <https://doi.org/10.1080/1068316X.2014.902457>
- Guéguen, N., Martin, A., Silone, F., & Pascual, A. (2016). The foot-in-the-door technique, crime, and the responsive bystander: A field experiment. *Crime Prevention and Community Safety*, 18(1), 60-68. <https://doi.org/10.1057/cpcs.2015.20>
- Guglielmo, S. (2015). Moral judgment as information processing: An integrative review. *Frontiers in Psychology*, 6:1637. <https://doi.org/10.3389/fpsyg.2015.01637>
- Guy, S., Kashima, Y., Walker, I., & O'Neill, S. (2014). Investigating the effects of knowledge and ideology on climate change beliefs. *European Journal of Social Psychology*, 44(5), 421-429. <https://doi.org/10.1002/ejsp.2039>
- Haidt, J. (2001). The emotional dog and its rational tail. *Psychological Review*, 108(4), 814-834. <https://doi.org/10.1037/0033-295x.108.4.814>
- Haidt, J. (2006). *The happiness hypothesis: Putting ancient wisdom and philosophy to the test of modern science*. Arrow Books.
- Haidt, J. (2011a, February 11). The bright future of post-partisan social psychology. *Edge*. Retrieved from [https://www.edge.org/conversation/jonathan\\_haidt-the-bright-future-of-post-partisan-social-psychology](https://www.edge.org/conversation/jonathan_haidt-the-bright-future-of-post-partisan-social-psychology)
- Haidt, J. (2011b). Out-take from The Righteous Mind: Virtue ethics. Retrived March 20, 2019, from <https://www.righteousmind.com/wp-content/uploads/2012/08/Righteous-Mind-outtake.virtue-ethics.pdf>
- Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Penguin Books.
- Haidt, J. (2016). The ethics of globalism, nationalism, and patriotism. *Minding Nature*, 9(3), 18-24. Retrieved from <https://www.humansandnature.org/the-ethics-of-globalism-nationalism-and-patriotism>
- Haim, M., Graefe, A., & Brosius, H. (2017). Burst of the filter bubble? *Digital Journalism*, 6(3), 330-343. <https://doi.org/10.1080/21670811.2017.1338145>
- Haji, I. (1997). An epistemic dimension of blameworthiness. *Philosophy and Phenomenological Research*, 57(3), 523-544. <https://doi.org/10.2307/2953747>

- Haksar, V. (1998). Moral agents. *The Routledge Encyclopedia of Philosophy*. Taylor and Francis. Retrieved January 5, 2019, from <https://www.rep.routledge.com/articles/thematic/moral-agents/v-1>
- Hall, M. P., & Raimi, K. T. (2018). Is belief superiority justified by superior knowledge? *Journal of Experimental Social Psychology*, 76, 290-306. <https://doi.org/10.1016/j.jesp.2018.03.001>
- Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology*, 39(4), 578-589. <https://doi.org/10.1037/0022-3514.39.4.578>
- Hampton, K. N., Shin, I., & Lu, W. (2017). Social media and political discussion: When online presence silences offline conversation. *Information, Communication & Society*, 20(7), 1090-1107. <https://doi.org/10.1080/1369118X.2016.1218526>
- Hanitzsch, T., van Dalen, A., & Steindl, N. (2018). Caught in the nexus: A comparative and longitudinal analysis of public trust in the press. *The International Journal of Press/Politics*, 23(1), 3-23. <https://doi.org/10.1177/1940161217740695>
- Hannikainen, I. R. (2018). Ideology between the lines. *Social Psychological and Personality Science*, OnlineFirst publication. <https://doi.org/10.1177/1948550618790230>
- Hansson, S. O. (2017). Science and pseudo-science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2017 Edition. Retrieved from <https://plato.stanford.edu/archives/sum2017/entries/pseudo-science/>
- Hardin, G. (1968). The tragedy of the commons. *Science*, 162(3859), 1243-1248. <https://doi.org/10.1126/science.162.3859.1243>
- Harman, E. (2011). Does moral ignorance exculpate? *Ratio*, 24(4), 443-468. <https://doi.org/10.1111/j.1467-9329.2011.00511.x>
- Harman, E. (2015). The irrelevance of moral uncertainty. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Vol. 10 (pp. 53-79). Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198738695.003.0003>
- Harmon-Jones, E., Harmon-Jones, C., & Levy, N. (2015). An action-based model of cognitive dissonance processes. *Current Directions in Psychological Science*, 24(3), 184-189. <https://doi.org/10.1177/0963721414566449>
- Harris, S. (2006, March 19). *Killing the Buddha* [post on author's homepage]. Retrieved May 9, 2019, from <https://samharris.org/killing-the-buddha/>
- Harris, S. (2010). *The moral landscape: How science can determine human values*. Bantam Press.
- Harris, S. (2012). *Free will*. Free Press.
- Harris, S. (2014, February 12). *The marionette's lament: A response to Daniel Dennett* [post on author's homepage]. Retrieved from <https://samharris.org/the-marionettes-lament/>
- Harris, S. (2014). *Waking up: A guide to spirituality without religion*. Simon & Schuster.
- Harris, T. (2017, July). *The need for a new design ethics* [Video and written statement]. Retrieved May 14, 2019, from <http://www.tristanharris.com/the-need-for-a-new-design-ethics/>
- Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4), 555-588. <https://doi.org/10.1037/a0015701>
- Hawkins, S., Yudkin, D., Juan-Torres, M., & Dixon, T. (2018). *Hidden tribes: A study of America's polarized landscape*. More in Common. Retrieved May 13, 2019, from <https://www.moreincommon.com/hidden-tribes>
- Heine, S. J., Lehman, D. R., Ide, E., Leung, C., Kitayama, S., Takata, T., & Matsumoto, H. (2001). Divergent consequences of success and failure in Japan and North America: An investigation of self-improving motivations and malleable selves. *Journal of Personality and Social Psychology*, 81(4), 599-615. <https://doi.org/10.1037/0022-3514.81.4.599>

- Heine, S. J., Takemoto, T., Moskaleiko, S., Lasaleta, J., & Henrich, J. (2008). Mirrors in the head: Cultural variation in objective self-awareness. *Personality and Social Psychology Bulletin*, 34(7), 879-887. <https://doi.org/10.1177/0146167208316921>
- Heinzelmann, N. (2018). Deontology defended. *Synthese*, 195(5), 5197-5216. <https://doi.org/10.1007/s11229-018-1762-3>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83. <https://doi.org/10.1017/S0140525X0999152X>
- Herman, B. (1993). *The practice of moral judgment*. Cambridge, MA: Harvard University Press.
- Hertwig, R., & Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, 11(3), 359-372. <https://doi.org/10.1177/1745691616635594>
- Hertzog, C. (2004). Metacognition in older adults: Implications for application. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied Metacognition* (pp. 169-196). Cambridge University Press. <https://doi.org/10.1017/cbo9780511489976.009>
- Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, 20(4), 311-335. <https://doi.org/10.1002/ejsp.2420200404>
- Heying, H., & Weinstein, B. (2017, December). Bonfire of the academies: Two professors on how leftist intolerance is killing higher education. *Washington Examiner*. Retrieved from <http://www.washingtonexaminer.com/bonfire-of-the-academies-two-professors-on-how-leftist-intolerance-is-killing-higher-education/article/2642973>
- Hibbing, J. R., Smith, K. B., & Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *The Behavioral and Brain Sciences*, 37(3), 297-307. <https://doi.org/10.1017/s0140525x13001192>
- Hmielowski, J. D., Feldman, L., Myers, T. A., Leiserowitz, A., & Maibach, E. (2014). An attack on science? Media use, trust in scientists, and perceptions of global warming. *Public Understanding of Science*, 23(7), 866-883. <https://doi.org/10.1177/0963662513480091>
- Hobbes, T., & Curley, E. (Ed.). (1994). *Leviathan*. Indianapolis, IN: Hackett Publishing Company.
- Hoffman, C. P., & Lutz, C. (2017). Spiral of silence 2.0: Political self-censorship among young Facebook users. *International Conference on Social Media & Society*. <https://doi.org/10.1145/3097286.3097296>
- Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioural therapy: A review of meta-analyses. *Cognitive Therapy and Research*, 36(5), 427-440. <https://doi.org/10.1007/s10608-012-9476-1>
- Honkela, T. (2017). *Rauhankone: Tekoälytutkijan testamentti* [The peace machine: The testament of an artificial intelligence researcher]. Gaudeamus.
- Hooper, N., Erdogan, A., Keen, G., Lawton, K., & McHugh, L. (2015). Perspective taking reduces the fundamental attribution error. *Journal of Contextual Behavioral Science*, 4(2), 69-72. <https://doi.org/10.1016/j.jcbs.2015.02.002>
- Hopthrow, T., Hooper, N., Mahmood, L., Meier, B. P., & Weger, U. (2017). Mindfulness reduces the correspondence bias. *The Quarterly Journal of Experimental Psychology*, 70(3), 351-360. <https://doi.org/10.1080/17470218.2016.1149498>
- Hornsey, M. J., & Fielding, K. S. (2017). Attitude roots and Jiu Jitsu persuasion: Understanding and overcoming the motivated rejection of science. *The American Psychologist*, 72(5), 459-473. <https://doi.org/10.1037/a0040437>
- Hornsey, M. J., Harris, E. A., Bain, P. G., & Fielding, K. S. (2016). Meta-analyses of the determinants and outcomes of belief in climate change. *Nature Climate Change*, 6, 622-626. <https://doi.org/10.1038/nclimate2943>
- Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018a). Relationships among conspiratorial beliefs, conservatism and climate scepticism across nations. *Nature Climate Change*, 8, 614-620. <https://doi.org/10.1038/s41558-018-0157-2>

- Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018b). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. *Health Psychology, 37*(4), 307-315. <https://doi.org/10.1037/hea0000586>
- Hortensius, R., & de Gelder, B. (2018). From empathy to apathy: The bystander effect revisited. *Current Directions in Psychological Science, 27*(4), 249-256. <https://doi.org/10.1177/0963721417749653>
- Howard, J., & Reiss, D. R. (2018). The anti-vaccine movement: A litany of fallacy and errors. In A. B. Kaufman & J. C. Kaufman (Eds.), *Pseudoscience: The Conspiracy Against Science* (pp. 195-219). Cambridge, MA: MIT Press.
- Howell, L. (2013). Digital wildfires in a hyperconnected world. WEF Report 2013. Retrieved February 3, 2019, from <http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/>
- Hume, D. (1748/1777). *An inquiry concerning human understanding*. P. Millican (Ed.). Hume Texts Online. Retrieved May 14, 2019, from on <https://davidhume.org/texts/e/>
- Hume, D. (1739–40). *A treatise of human nature*. A. Merivale, & P. Millican (Eds.). Hume Texts Online. Retrieved February 5, 2019, from <https://davidhume.org/texts/t/full>
- Hursthouse, R., & Pettigrove, G. (2018). Virtue ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2018 Edition. Retrieved from <https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/>
- Huskey, R., Bowman, N., Eden, A., Grizzard, M., Hahn, L., Lewis, R., Matthews, N., Tamborini, R., Walther, J. B., & Weber, R. (2018). Things we know about media and morality. *Nature Human Behaviour, 2*, 315. <https://doi.org/10.1038/s41562-018-0349-9>
- Inbar, Y., & Lammers, J. (2012). Political diversity in social and personality psychology. *Perspective on Psychological Science, 7*(5), 496-503. <https://doi.org/10.1177/1745691612448792>
- Inbar, Y., Pizarro, D. A., & Bloom, P. (2012a). Disgusting smells cause decreased liking of gay men. *Emotion, 12*(1), 23-27. <https://doi.org/10.1037/a0023984>
- Inbar, Y., Pizarro, D., Iyer, R., & Haidt, J. (2012b). Disgust sensitivity, political conservatism, and voting. *Social Psychological and Personality Science, 3*(5), 537-544. <https://doi.org/10.1177/1948550611429024>
- Internet World Stats. (2017). Facebook users in the world. Retrieved December 13, 2018, from <https://www.internetworldstats.com/facebook.htm>
- Internet World Stats. (2019a). Internet growth statistics. Retrieved May 8, 2019, from <https://www.internetworldstats.com/emarketing.htm>
- Internet World Stats. (2019b). Internet usage statistics. Retrieved May 8, 2019, from <https://www.internetworldstats.com/stats.htm>
- IPCC (Intergovernmental Panel on Climate Change). (2014). *Climate change 2014: Synthesis report. Contributions of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Retrieved from [http://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR\\_AR5\\_FINAL\\_full.pdf](http://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_FINAL_full.pdf)
- IPCC (Intergovernmental Panel on Climate Change). (2018). *Global warming of 1.5 °C – An IPCC special report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. Retrieved from <http://www.ipcc.ch/report/sr15/>
- Iyengar, S., & Krupenkin, M. (2018). The strengthening of partisan affect. *Political Psychology, 39*(S1), 201-218. <https://doi.org/10.1111/pops.12487>
- Iyengar, S., & Massey, D. S. (2019). Scientific communication in a post-truth society. *PNAS, 116*(16), 7656-7661. <https://doi.org/10.1073/pnas.1805868115>
- Jamieson, D. (2007). When utilitarians should be virtue theorists. *Utilitas, 19*(2), 160-183. <https://doi.org/10.1017/S0953820807002452>
- Jamieson, K. H., Kahan, D., & Scheufele, D. A. (Eds.). (2017). *The Oxford handbook of the science of science communication*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190497620.001.0001>

- Johnson, A. J., Lee, S. K., Cionea, I. A., & Massey, Z. B. (2018). The benefits and challenges of new media for intercultural conflict. In N. Bilge & M. I. Marino (Eds.), *Reconceptualizing New Media and Intercultural Communication in a Networked Society* (pp. 171-197). Hershey, PA: IGI Global. <https://doi.org/10.4018/978-1-5225-3784-7.ch007>
- Jones, D. T. (2015, September 6). Pyrrho and the Buddha: Reasons to be sceptical. *Western Buddhist Review*. Retrieved from <https://thebuddhistcentre.com/westernbuddhistreview/pyrrho-and-buddha-reasons-be-sceptical>
- Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, 530, 473-476. <https://doi.org/10.1038/nature16981>
- Joslyn, M. R., & Haider-Markel, D. P. (2014). Who knows best? Education, partisanship, and contested facts. *Politics & Policy*, 42(6), 919-947. <https://doi.org/10.1111/polp.12098>
- Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political Psychology*, 38(2), 167-208. <https://doi.org/10.1111/pops.12407>
- Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision Making*, 8(4), 407-424. <https://doi.org/10.2139/ssrn.2182588>
- Kahan, D. M. (2015). Climate-science communication and the measurement problem. *Advances in Political Psychology*, 36(S1), 1-43. <https://doi.org/10.1111/pops.12244>
- Kahan, D. M. (2017a). 'Ordinary science intelligence': A science-comprehension measure for study of risk and science communication, with notes on evolution and climate change. *Journal of Risk Research*, 20(8), 995-1016. <https://doi.org/10.1080/13669877.2016.1148067>
- Kahan, D. M. (2017b). Misconceptions, misinformation, and the logic of identity-protective cognition. Cultural Cognition Project Working Paper Series No. 164; Yale Law School, Public Law Research Paper No. 605; Yale Law & Economics Research Paper No. 575. <https://doi.org/10.2139/ssrn.2973067>
- Kahan, D. M. (2018, December). Why smart people are vulnerable to putting tribe before truth. *Scientific American*. Retrieved from <https://blogs.scientificamerican.com/observations/why-smart-people-are-vulnerable-to-putting-tribe-before-truth/>
- Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, 14(2), 147-174. <https://doi.org/10.1080/13669877.2010.511246>
- Kahan, D. M., & Peters, E. (2017). Rumors of the 'nonreplication' of the 'motivated numeracy effect' are greatly exaggerated. Yale Law & Economics Research Paper No. 584. <https://doi.org/10.2139/ssrn.3026941>
- Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Quellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, 2, 732-735. <https://doi.org/10.1038/nclimate1547>
- Kahan, D. M., Landrum, A., Carpenter, K., Helft, L., & Jamieson, K. H. (2017a). Science curiosity and political information processing. *Advances in Political Psychology*, 38(S1), 179-199. <https://doi.org/10.1111/pops.12396>
- Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017b). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, 1(1), 54-86. <https://doi.org/10.1017/bpp.2016.2>
- Kahneman, D. (2011). *Thinking, fast and slow*. Penguin Books.
- Kanai, R., Feilden, T., Firth, C., & Rees, G. (2011). Political orientations are correlated with brain structure in young adults. *Current Biology*, 21(8), 677-680. <https://doi.org/10.1016/j.cub.2011.03.017>
- Kane, R. (1998). *The significance of free will*. New York: Oxford University Press.
- Kane, R. (2002). Responsibility, reactive attitudes and free will: Reflections on Wallace's theory [Review]. *Philosophy and Phenomenological Research*, 64(3), 693-698. <https://doi.org/10.1111/j.1933-1592.2002.tb00175.x>
- Kant, I. (1964). *Groundwork of the metaphysics of morals* (H. J. Paton, Trans.). New York: Harper & Row.

- Karlamangla, S. (2019, March). Anti-vaccine activists have doctors ‘terrorized into silence’ with online harassment. *Los Angeles Times*. Retrieved from <https://www.latimes.com/local/california/la-me-ln-vaccine-attacks-20190317-story.html>
- Kaufman, A. B., & Kaufman, J. C. (Eds.). (2018). *Pseudoscience: The conspiracy against science*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/10747.001.0001>
- Kauppinen, A. (2013). A Humean theory of moral intuition. *Canadian Journal of Philosophy*, 43(3), 360-381. <https://doi.org/10.1080/00455091.2013.857136>
- Kerber, K. W. (1984). The perception of nonemergency helping situations: Costs, rewards, and the altruistic personality. *Journal of Personality*, 52(2), 177-187. <https://doi.org/10.1111/j.1467-6494.1984.tb00351.x>
- Khan, S. (2011, March). *Let's use video to reinvent education* [Video]. Retrieved from [https://www.ted.com/talks/salman\\_khan\\_let\\_s\\_use\\_video\\_to\\_reinvent\\_education](https://www.ted.com/talks/salman_khan_let_s_use_video_to_reinvent_education)
- Khoury, B., Lecomte, T., Fortin, G., Masse, M., Therien, P., Bouchard, V., Chapleau, M.-A., Paquin, K., & Hofmann, S. G. (2013). Mindfulness-based therapy: A comprehensive meta-analysis. *Clinical Psychology Review*, 33(6), 763-771. <https://doi.org/10.1016/j.cpr.2013.05.005>
- Khoury, B., Sharma, M., Rush, S. E., & Fournier, C. (2015). *Journal of Psychosomatic Research*, 78(6), 519-528. <https://doi.org/10.1016/j.jpsychores.2015.03.009>
- Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., & Theobalt, C. (2018). Deep Video Portraits. *ACM Transactions on Graphics*, 37(4), 163. <https://doi.org/10.1145/3197517.3201283>
- King, M. (2017). Tracing the epistemic condition. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition* (pp. 266-280). Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.003.0015>
- Kirby, J. N. (2017). Compassion interventions: The programmes, the evidence, and implications for research and practice. *Psychology and Psychotherapy*, 90(3), 432-455. <https://doi.org/10.1111/papt.12104>
- Kirby, J. N., Tellegen, C. L., & Steindl, S. R. (2017). A meta-analysis of compassion-based interventions: Current state of knowledge and future directions. *Behavior Therapy*, 48(6), 778-792. <https://doi.org/10.1016/j.beth.2017.06.003>
- Kirchner, A., Boiger, M., Uchida, Y., Norasakkunkit, V., Verduyn, P., & Mesquita, B. (2018). Humiliated fury is not universal: The co-occurrence of anger and shame in the United States and Japan. *Cognition and Emotion*, 32(6), 1317-1328. <https://doi.org/10.1080/02699931.2017.1414686>
- Kitayama, S., Ishii, K., Imada, T., Takemura, K., & Ramaswamy, J. (2006a). Voluntary settlement and the spirit of independence: Evidence from Japan's "Northern frontier". *Journal of Personality and Social Psychology*, 91(3), 369-384. <https://doi.org/10.1037/0022-3514.91.3.369>
- Kitayama, S., Mesquita, B., & Karasawa, M. (2006b). Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States. *Journal of Personality and Social Psychology*, 91(5), 890-903. <https://doi.org/10.1037/0022-3514.91.5.890>
- Klimecki, O., Ricard, M., & Singer, T. (2013). Empathy versus compassion – Lessons from 1<sup>st</sup> and 3<sup>rd</sup> person methods. In T. Singer & M. Bolz (Eds.), *Compassion – Bridging Practice and Science* (pp. 272-287). Max Planck Society, Munich, Germany. Retrieved from <http://www.compassion-training.org/?page=download&lang=en>
- Klimecki, O. M., Leiberg, S., Ricard, M., & Singer, T. (2014). Differential pattern of functional brain plasticity after compassion and empathy training. *Social Cognitive and Affective Neuroscience*, 9(6), 873-879. <https://doi.org/10.1093/scan/nst060>
- Knickelbine, M. (2011, October). *The ethics of impermanence* [Blog post]. Secular Buddhist Association. Retrieved from <https://secularbuddhism.org/the-ethics-of-impermanence/>
- Knobe, J., & Doris, J. M. (2010). Responsibility. In J. M. Doris & the Moral Psychology Research Group (Eds.), *The Moral Psychology Handbook* (pp. 321-354). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199582143.003.0011>

- Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2017). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research*, OnlineFirst publication. <https://doi.org/10.1177/0093650217719596>
- Koller, D. (2012, June). *What we're learning from online education* [Video]. Retrieved from [https://www.ted.com/talks/daphne\\_koller\\_what\\_we\\_re\\_learning\\_from\\_online\\_education](https://www.ted.com/talks/daphne_koller_what_we_re_learning_from_online_education)
- Konishi, N., Oe, T., Shimizu, H., Tanaka, K., & Ohtsubo, Y. (2017). Perceived shared condemnation intensifies punitive moral emotions. *Scientific Reports*, 7, 7289. <https://doi.org/10.1038/s41598-017-07916-z>
- Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), 34-53. <https://doi.org/10.1037/0278-7393.27.1.34>
- Korsgaard, C. (1996a). *Creating the kingdom of ends*. Cambridge: Cambridge University Press.
- Korsgaard, C. (1996b). *The sources of normativity*. Cambridge: Cambridge University Press.
- Kozak, A. (2018). Understanding pseudoscience vulnerability through epistemological development, critical thinking, and science literacy. In A. B. Kaufman & J. C. Kaufman (Eds.), *Pseudoscience: The Conspiracy Against Science* (pp. 223-238). Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/9780262037426.003.0009>
- Kraut, R. (2018). Aristotle's ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2018 Edition. Retrieved from <https://plato.stanford.edu/archives/sum2018/entries/aristotle-ethics/>
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Krumrei-Mancuso, E. J., Haggard, M. C., LaBouff, J. P., & Rowatt, W. C. (2019). Links between intellectual humility and acquiring knowledge. *The Journal of Positive Psychology*, 1-16. <https://doi.org/10.1080/17439760.2019.1579359>
- Kumar, M.-L. (The Enthusiastic Buddhist). (2013a, October 5). *Buddhist beliefs: The four noble truths* [Video]. Retrieved from <https://www.youtube.com/watch?v=zerG3watL0k>
- Kumar, M.-L. (The Enthusiastic Buddhist). (2013b, October 20). *Buddhist teachings: The noble eightfold path* [Video]. Retrieved from <https://www.youtube.com/watch?v=pPk-pxhyYeg>
- Kumar, M.-L. (The Enthusiastic Buddhist). (2014, July 20). *The nature of mind, five defilements & three poisons in Buddhism* [Video]. Retrieved from [https://www.youtube.com/watch?v=mk\\_2xRcFIU0](https://www.youtube.com/watch?v=mk_2xRcFIU0)
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Kuran, T. (1997). *Private truths, public lies – The social consequences of preference falsification*. Harvard University Press.
- Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, 28(2), 75-84. <https://doi.org/10.1016/j.evolhumbehav.2006.06.001>
- Kuzminski, A. (2008). *Pyrrhonism: How the ancient Greeks reinvented Buddhism*. Lexington Books.
- Landry, N., Gifford, R., Milfont, T. L., Weeks, A., & Arnocky, A. (2018). Learned helplessness moderates the relationship between environmental concern and behavior. *Journal of Environmental Psychology*, 55, 18-22. <https://doi.org/10.1016/j.jenvp.2017.12.003>
- Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*, 10(3), 215-221. <https://doi.org/10.1037/h0026570>
- Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York, NY: Appleton-Century-Croft.

- Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, 89(2), 308-324. <https://doi.org/10.1037/0033-2909.89.2.308>
- Latané, B., & Rodin, J. (1969). A lady in distress: Inhibiting effects of friends and strangers on bystander intervention. *Journal of Experimental Social Psychology*, 5(2), 189-202. [https://doi.org/10.1016/0022-1031\(69\)90046-8](https://doi.org/10.1016/0022-1031(69)90046-8)
- Lavin, D. (2008). The second person standpoint: Morality, respect, and accountability [Review]. *Notre Dame Philosophical Reviews*, 2008.01.05. Retrieved from <https://ndpr.nd.edu/news/the-second-person-standpoint-morality-respect-and-accountability/>
- Lerman, K., Yan, X., & Wu, X.-Z. (2016). The “majority illusion” in social networks. *PLOS ONE*, 11(2), e0147617. <https://doi.org/10.1371/journal.pone.0147617>
- Levine, R. V. (2003). The kindness of strangers. *American Scientist*, 91(3), 226-233. Retrieved from <https://www.jstor.org/stable/27858212>
- Levine, M., Cassidy, C., Brazier, G., & Reicher, S. (2002). Self-categorization and bystander non-intervention: Two experimental studies. *Journal of Applied Social Psychology*, 32(7), 1452-1463. <https://doi.org/10.1111/j.1559-1816.2002.tb01446.x>
- Levine, M., & Crowther, S. (2008). The responsive bystander: How social group membership and group size can encourage as well as inhibit bystander intervention. *Journal of Personality and Social Psychology*, 95(6), 1429-1439. <https://doi.org/10.1037/a0012634>
- Levine, R. V., Martinez, T. S., Brase, G., & Sorenson, K. (1994). Helping in 36 U.S. cities. *Journal of Personality and Social Psychology*, 67(1), 69-82. <https://doi.org/10.1037/0022-3514.67.1.69>
- Levine, R. V., Norenzayan, A., & Philbrick, K. (2001). Cross-cultural differences in helping strangers. *Journal of Cross-Cultural Psychology*, 32(5), 543-560. <https://doi.org/10.1177/0022022101032005002>
- Levine, R. V., Reysen, S., & Ganz, E. (2008). The kindness of strangers revisited: A comparison of 24 US cities. *Social Indicators Research*, 85(3), 461-481. <https://doi.org/10.1007/s11205-007-9091-9>
- Levy, N. (2005). The good, the bad, and the blameworthy. *Journal of Ethics and Social Psychology*, 1(2), 2-16. <https://doi.org/10.26556/jesp.v1i2.6>
- Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199601387.001.0001>
- Levy, N. (2014). *Consciousness and moral responsibility*. Oxford: Oxford University Press.
- Levy, N. (2017). Methodological conservatism and the epistemic condition. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition* (pp. 252-265). Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.003.0014>
- Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353-369. <https://doi.org/10.1016/j.jarmac.2017.07.008>
- Lewandowsky, S., Gignac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, 3, 399-404. <https://doi.org/10.1038/nclimate1720>
- Lewandowsky, S., & Oberauer, K. (2016). Motivated rejection of science. *Current Directions in Psychological Science*, 25(4), 217-222. <https://doi.org/10.1177/0963721416654436>
- Lewis, G. B., Palm, R., & Beng, B. (2018). Cross-national variation in determinants of climate change concern. *Environmental Politics*. <https://doi.org/10.1080/09644016.2018.1512261>
- Lewis, P. (2017, November). ‘I see things differently’: James Damore on his autism and the Google memo. *The Guardian*. Retrieved from <https://www.theguardian.com/technology/2017/nov/16/james-damore-google-memo-interview-autism-regrets>

- Lim, D., Condon, P., & DeSteno, D. (2015). Mindfulness and compassion: An examination of mechanism and scalability. *PLOS ONE*, 10(2), e018221. <https://doi.org/10.1371/journal.pone.0118221>
- Linden, S. van der, Leiserowitz, A., & Maibach, E. (2017). Scientific agreement can neutralize politicization of facts. *Nature Human Behaviour*, 2, 2-3. <https://doi.org/10.1038/s41562-017-0259-2>
- Lobato, E. J. C., & Zimmerman, C. (2018). The psychology of (pseudo)science: Cognitive, social, and cultural factors. In A. B. Kaufman & J. C. Kaufman (Eds.), *Pseudoscience: The Conspiracy Against Science* (pp. 21-43). Cambridge, MA: MIT Press.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098-2109. <https://doi.org/10.1037/0022-3514.37.11.2098>
- Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, 10:1759, 1-9. <https://doi.org/10.1038/s41467-019-09311-w>
- Loury, G. C. (1994). Self-Censorship in public discourse – A theory of “political correctness” and related phenomena. *Rationality and Society*, 6(4), 428-461. <https://doi.org/10.1177/1043463194006004002>
- Luskin, R. C., O’Flynn, I., Fishkin, J. S., & Russell, D. (2014). Deliberating across deep divides. *Political Studies*, 62(1), 116-135. <https://doi.org/10.1111/j.1467-9248.2012.01005.x>
- MacIntyre, A. (1957). Determinism. *Mind*, 66(261), 28-41. <https://doi.org/10.1093/mind/LXVI.261.28>
- MacIntyre, A. (1981/1984/2007). *After virtue*. University of Notre Dame Press.
- Madrigal, A. C. (2018, September). India’s lynching epidemic and the problem with blaming tech. *The Atlantic*. Retrieved from <https://www.theatlantic.com/technology/archive/2018/09/whatsapp/571276/>
- Maheshwari, S. (2017, November). On YouTube Kids, startling videos slip past filters. *The New York Times*. Retrieved from <https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html>
- Mahmoodi, A., Band, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C. D., Roepstorff, A., Rees, G., & Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *PNAS*, 112(12), 3835-3840. <https://doi.org/10.1073/pnas.1421692112>
- Maio, G. R., Haddock, G., & Verplanken, B. (Eds.). (2019). *The psychology of attitudes & attitude change* (3<sup>rd</sup> edition). SAGE.
- Malka, A., Lelkes, Y., Srivastava, S., Cohen, A. B., & Miller, D. T. (2012). The association of religiosity and political conservatism: The role of political engagement. *Political Psychology*, 33(2), 275-299. <https://doi.org/10.1111/j.1467-9221.2012.00875.x>
- Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895-919. <https://doi.org/10.1037/0033-2909.132.6.895>
- Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, 25(2), 147–186. <https://doi.org/10.1080/1047840x.2014.877340>
- Mann, M. E. (2016, December). I’m a scientist who has gotten death threats. I fear what may happen under Trump. *The Washington Post*. Retrieved from [https://www.washingtonpost.com/opinions/this-is-what-the-coming-attack-on-climate-science-could-look-like/2016/12/16/e015cc24-bd8c-11e6-94ac-3d324840106c\\_story.html](https://www.washingtonpost.com/opinions/this-is-what-the-coming-attack-on-climate-science-could-look-like/2016/12/16/e015cc24-bd8c-11e6-94ac-3d324840106c_story.html)
- Marcus Aurelius. (2003). *Meditations* (G. Hays, Trans.). New York: Modern Library.
- Markey, P. M. (2000). Bystander intervention in computer-mediated communication. *Computers in Human Behavior*, 16(2), 183-188. [https://doi.org/10.1016/S0747-5632\(99\)00056-4](https://doi.org/10.1016/S0747-5632(99)00056-4)
- Markowitz, E. M., & Shariff, A. F. (2012). Climate change and moral judgement. *Nature Climate Change*, 2, 243-247. <https://doi.org/10.1038/nclimate1378>

- Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102(1), 72-90. <https://doi.org/10.1037/0033-2909.102.1.72>
- Martin, N. D., Rigoni, D., & Vohs, K. D. (2017). Free will beliefs predict attitudes toward unethical behaviour and criminal punishment. *PNAS*, 114(28), 7325-7330. <https://doi.org/10.1073/pnas.1702119114>
- Martinez, M. (2018, November). Burned to death because of a rumour on WhatsApp. *BBC*. Retrieved from <https://www.bbc.com/news/world-latin-america-46145986>
- Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81(5), 922-934. <https://doi.org/10.1037/0022-3514.81.5.922>
- Matthes, J., Knoll, J., & von Sikorski, C. (2018). The "spiral of silence" revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression. *Communication Research*, 45(1), 3-33. <https://doi.org/10.1177/0093650217745429>
- McCoy, C. (2018). The social characteristics of Americans opposed to vaccination: Beliefs about vaccine safety versus views of U.S. vaccination policy. *Critical Public Health*, 1-12. <https://doi.org/10.1080/09581596.2018.1501467>
- McGrath, A. (2017). Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass*, 11(12), e12362. <https://doi.org/10.1111/spc3.12362>
- McGraw, A. P., & Warren, C. (2010). Benign violations – Making immoral behaviour funny. *Psychological Science*, 21(8), 1141-1149. <https://doi.org/10.1177/0956797610376073>
- McGraw, A. P., Warren, C., Williams, L. E., & Leonard, B. (2012). Too close for comfort, or too far to care? Finding humor in distant tragedies and close mishaps. *Psychological Science*, 23(10), 1215-1223. <https://doi.org/10.1177/0956797612443831>
- McHugh, C. (2013). Epistemic responsibility and doxastic agency. *Philosophical Issues*, 23(1), 132-157. <https://doi.org/10.1111/phils.12007>
- McIntosh, R. D., Fowler, E. A., Lyu, T., & Della Sala, S. (2019). Wise up: Clarifying the role of metacognition in the Dunning-Kruger effect. *Journal of Experimental Psychology: General*. Advance online publication. <https://doi.org/10.1037/xge0000579>
- McKay, R., & Whitehouse, H. (2015). Religion and morality. *Psychological Bulletin*, 141(2), 447-473. <https://doi.org/10.1037/a0038455>
- McKenna, M. (2008). Putting the lie on the control condition for moral responsibility. *Philosophical Studies*, 139(1), 29-37. <https://doi.org/10.1007/s11098-007-9100-5>
- McKenna, M., & Coates, D. J. (2018). Compatibilism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2018 Edition. Retrieved from <https://plato.stanford.edu/archives/fall2018/entries/compatibilism/>
- McRae, E. (2015). Metabolizing anger: A tantric Buddhist solution to the problem of moral anger. *Philosophy East and West*, 65(2), 466-484. <https://doi.org/10.1353/pew.2015.0041>
- Mele, A. (1995). *Autonomous agents*. Oxford: Oxford University Press.
- Metz, T., & Gaie, J. B. R. (2010). The African ethic of *ubuntu/botho*: Implications for research on morality. *Journal of Moral Education*, 39(3), 273-290. <https://doi.org/10.1080/03057240.2010.497609>
- Mildenberger, M., Marlon, J., Howe, P., & Leiserowitz, A. (2015). The spatial distribution of Republic and Democratic climate opinions at state and local scales. *Climatic Change*, 145(3-4), 539-548. <https://doi.org/10.1007/s10584-017-2103-0>
- Milfont, T. L., Wilson, M. S., & Sibley, C. G. (2017). The public's belief in climate change and its human cause are increasing over time. *PLOS ONE*, 12(3), e0174246. <https://doi.org/10.1371/journal.pone.0174246>
- Miller, J. D. (1998). The measurement of civic scientific literacy. *Public Understanding of Science*, 7(3), 203-223. <https://doi.org/10.1088/0963-6625/7/3/001>

- Miller, J. M., Saunders, K. L., & Farhart, C. E. (2015). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. *American Journal of Political Science*, 60(4), 824-844. <https://doi.org/10.1111/ajps.12234>
- Miller, S. (2017). Ignorance, technology, and collective responsibility. In R. Peels (Ed.), *Perspectives on Ignorance from Moral and Social Philosophy* (chapter 12). Routledge.
- Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, 116(2), 215-236. <https://doi.org/10.1037/pspa0000137>
- Montmarquet, J. A. (1999). Zimmerman on culpable ignorance. *Ethics*, 109(4), 842-845. <https://doi.org/10.1086/233949>
- Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, 2, 389-396. <https://doi.org/10.1038/s41562-018-0353-0>
- Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502-517. <https://doi.org/10.1037/0033-295X.115.2.502>
- Moriarty, T. (1975). Crime, commitment, and the responsive bystander: Two field experiments. *Journal of Personality and Social Psychology*, 31(2), 370-376. <https://doi.org/10.1037/h0076288>
- Motta, M., Callaghan, T., & Sylvester, S. (2018). Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Social Science & Medicine*, 211, 274-281. <https://doi.org/10.1016/j.socscimed.2018.06.032>
- Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7), 959-977. <https://doi.org/10.1080/1369118X.2018.1444076>
- Nagel, T. (1986). *The view from nowhere*. Oxford: Oxford University Press.
- Nagel, T. (2013, November). You can't learn about morality from brain scans. *The New Republic*. Retrieved from <https://newrepublic.com/article/115279/joshua-greenes-moral-tribes-reviewed-thomas-nagel>
- NASA (National Aeronautics and Space Administration). (2019, February). 2018 fourth warmest year in continued warming trend, according to NASA, NOAA [Press release 19-002]. New York: Goddard Institute for Space Studies. Retrieved from <https://www.nasa.gov/press-release/2018-fourth-warmest-year-in-continued-warming-trend-according-to-nasa-noaa>
- Nawaz, M. (2015, November 18). *Je suis Muslim: how universal secular rights protect Muslim communities the most* [Video]. Retrieved from <https://bigthink.com/videos/maajid-nawaz-on-islamic-reform>
- Nawaz, M. (2016a, February 29). *How to think critically about Islam without denigrating its practitioners* [Video]. Retrieved from <https://bigthink.com/videos/maajid-nawaz-on-sam-harris-and-political-correctness>
- Nawaz, M. (2016b, October). I'm a Muslim reformer. Why am I being smeared as an 'anti-Muslim extremist'? *Daily Beast*. Retrieved from <https://www.thedailybeast.com/im-a-muslim-reformer-why-am-i-being-smeared-as-an-anti-muslim-extremist>
- Nelkin, D. K. (2011a). *Making sense of freedom and responsibility*. Oxford University Press.
- Nelkin, D. K. (2011b). Sher, George. Who knew? Responsibility without awareness [Review]. *Ethics*, 121(3), 675-680. <https://doi.org/10.1086/659370>
- Nelkin, D. K., & Rickless, S. C. (2017). Moral responsibility for unwitting omissions: A new tracing view. In D. K. Nelkin, & S. C. Rickless (Eds.), *The Ethics and Law of Omissions* (pp. 106-130). Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780190683450.003.0006>
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1), 109-133. <https://doi.org/10.1037/0033-2909.95.1.109>
- Nelson, J. L., & Webster, J. G. (2017). The myth of partisan selective exposure: A portrait of the online political news audience. *Social Media + Society*, 3(3), 1-13. <https://doi.org/10.1177/2056305117729314>

- Neuronicus. (2018, January). *The FIRSTS: The Dunning–Kruger effect (1999) or the unskilled-and-unaware phenomenon* [Blog post]. Retrieved May 13, 2019, from <https://scientiportal.wordpress.com/2018/01/10/the-unskilled-and-unaware-phenomenon/>
- Newport, F., & Busteed, B. (2017, August). Why are Republicans down on higher ed? *Gallup*. Retrieved from <https://news.gallup.com/poll/216278/why-republicans-down-higher.aspx>
- Nguyen, C. T. (2011). *An ethics of uncertainty: Moral disagreement and moral humility* (Doctoral dissertation). Retrieved from <https://philpapers.org/rec/NGUAEO-2>
- Nguyen, C. T. (2018a, April). *Escape the echo chamber*. Aeon. Retrieved from <https://aeon.co/essays/why-its-as-hard-to-escape-an-echo-chamber-as-it-is-to-flee-a-cult>
- Nguyen, C. T. (2018b). Echo chambers and epistemic bubbles. *Episteme*, 1-21. Online publication. <https://doi.org/10.1017/epi.2018.32>
- Nguyen, C. T. (2018c). Cognitive islands and runaway echo chambers: Problems for epistemic dependence on experts. *Synthese*, 1-19. <https://doi.org/10.1007/s11229-018-1692-0>
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175-220. <https://doi.org/10.1037/1089-2680.2.2.175>
- Nir, L. (2011). Motivated reasoning and public opinion perception. *Public Opinion Quarterly*, 75(3), 504-532. <https://doi.org/10.1093/poq/nfq076>
- Nisbet, E. C., Cooper, K. E., & Garrett, R. K. (2015). The partisan brain: How dissonant science messages lead conservatives and liberals to (dis)trust science. *The ANNALS of the American Academy of Political and Social Science*, 658(1), 36-66. <https://doi.org/10.1177/0002716214555474>
- Nisbet, M. C. (2009). Communicating climate change: Why frames matter for public engagement. *Environment Magazine*, 51(2), 12-23. <https://doi.org/10.3200/ENVT.51.2.12-23>
- Nisbett, R. (2016). *Mindware: Tools for smart thinking*. London: Penguin Books.
- Norgaard, K. M. (2009). *Cognitive and behavioural challenges in responding to climate change*. The World Bank. <https://doi.org/10.1596/1813-9450-4940>
- Norris, P., & Inglehart, R., (2019). *Cultural backlash: Trump, Brexit, and authoritarian-populism*. New York: Cambridge University Press. <https://doi.org/10.1017/9781108595841>
- Novella, S. (2018). *The skeptics' guide to the universe: How to know what's really real in a world increasingly full of fake*. Hodder & Stoughton.
- Novella, S. (2019, January). *Dunning Kruger and GMO opposition* [Blog post]. Retrieved from <https://theness.com/neurologicablog/index.php/dunning-kruger-and-gmo-opposition/>
- Nozick, R. (1981). *Philosophical explanations*. Cambridge, MA: Harvard University Press.
- Oldways (Boston, MA). (2015). *Oldways common ground consensus statement on healthy eating*. Oldways Finding Common Ground conference, November 17–18. Retrieved May 14, 2019, from <https://oldwayspt.org/programs/oldways-common-ground/oldways-common-ground-consensus>
- Olson, E. T. (2017). Personal identity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2017 Edition. Retrieved from <https://plato.stanford.edu/archives/sum2017/entries/identity-personal/>
- Online Exchange on “Democratic Deconsolidation” (2017). *Journal of Democracy*. Retrieved from <https://www.journalofdemocracy.org/online-exchange-“democratic-deconsolidation”>
- Oreskes, N., & Conway, E. M. (2008). Challenging knowledge: How climate science became a victim of the Cold War. In Robert N. Proctor & Londa Schiebinger (Eds.), *Agnotology: The making and unmaking of ignorance* (pp. 55-89). California, USA: Stanford University Press.

- Oreskes, N. (2018). The scientific consensus on climate change: How do we know we're not wrong? In Elisabeth A. Lloyd & Eric Winsberg (Eds.), *Climate modelling – Philosophical and conceptual issues* (pp. 31-64). London, England: Palgrave Macmillan, Cham. [https://doi.org/10.1007/978-3-319-65058-6\\_2](https://doi.org/10.1007/978-3-319-65058-6_2)
- Orzel, C. (2013, November). *Interference with 10,000-particle "particles": "Matter-wave interference with particles selected from a molecular library with masses exceeding 10000 amu"* [Blog post]. Retrieved from <https://scienceblogs.com/principles/2013/11/12/interference-with-10000-particle-particles-matter-wave-interference-with-particles-selected-from-a-molecular-library-with-masses-exceeding-10000-amu>
- Pakaluk, M. (2005). *Aristotle's Nicomachean Ethics: An introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511802041>
- Pantin, H., & Carver, C. S. (1982). Induced competence and the bystander effect. *Journal of Applied Social Psychology*, 12(2), 100-111. <https://doi.org/10.1111/j.1559-1816.1982.tb00852.x>
- Parks, B. D. (2009). *Ultimate moral responsibility is impossible: A new defense of the basic argument*. Unpublished manuscript. Retrieved May 10, 2019, from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.395.3763>
- Peels, R. (2011). Tracing culpable ignorance. *Logos & Episteme*, 2(4), 575-582. <https://doi.org/10.5840/logos-episteme2011246>
- Peels, R. (2014). What kind of ignorance excuses? *Philosophical Quarterly*, 64(256), 478-496. <https://doi.org/10.1093/pq/pqu013>
- Pennycook, G., Bear, A., Collins, E., & Rand, D. G. (2019). The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. Working paper, Available at SSRN. <https://doi.org/10.2139/ssrn.3035384>
- Pennycook, G., Cannon, T., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865-1880. <https://doi.org/10.1037/xge0000465>
- Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015a). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, 10(6), 549-563.
- Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015b). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, 24(6), 425-432. <https://doi.org/10.1177/0963721415604610>
- Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015c). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, 80, 34-72. <https://doi.org/10.1016/j.cogpsych.2015.05.001>
- Pennycook, G., & Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. *PNAS*, 116(7), 2521-2526. <https://doi.org/10.1073/pnas.1806781116>
- Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39-50. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pennycook, G., & Rand, D. G. (2019c). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*. Preprint working paper, Available at SSRN. <https://doi.org/10.2139/ssrn.3023545>
- Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning-Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review*, 24(6), 1774-1784. <https://doi.org/10.3758/s13423-017-1242-7>
- Pereboom, D. (2013). Optimistic skepticism about free will. In P. Russell & O. Deery (Eds.), *The Philosophy of Free Will: Essential Readings from the Contemporary Debates* (pp. 421-449). Oxford University Press.
- Pereboom, D. (2015). A notion of moral responsibility immune to the threat from causal determination. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The Nature of Moral Responsibility: New Essays* (pp. 281-296). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199998074.003.0013>

- Pereboom, D., & Caruso, G. D. (2018). Hard-incompatibilist existentialism: Neuroscience, punishment, and meaning in life. In G. D. Caruso, & O. Flanagan (Eds.), *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience* (pp. 193-222). Oxford University Press.
- Perrino, N. (2017, June). Yale 2.0 at Evergreen State College? [Case summary]. *Foundation for Individual Rights in Education (FIRE)*. Retrieved from <https://www.thefire.org/yale-2-0-at-evergreen-state-college/>
- Peterson, C., & Seligman, M. (2004). *Character strengths and virtues: A handbook and classification*. Oxford University Press.
- Pew Research Center. (2014). *Political polarization in the American public*. Retrieved from <http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/>
- Pew Research Center. (2015). *Global concern about climate change, broad support for limiting emissions*. Retrieved from <http://www.pewglobal.org/2015/11/05/global-concern-about-climate-change-broad-support-for-limiting-emissions/>
- Pew Research Center. (2016). *Partisanship and political animosity in 2016*. Retrieved from <http://www.people-press.org/2016/06/22/partisanship-and-political-animosity-in-2016/>
- Pew Research Center. (2017a). *News use across social media platforms 2017*. Retrieved from <http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/>
- Pew Research Center. (2017b). *The partisan divide on political values grows even wider*. Retrieved from <http://www.people-press.org/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/>
- Pew Research Center. (2017c). *The future of truth and misinformation online*. Retrieved from <http://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/>
- Pew Research Center. (2017d). *How much does science knowledge influence people's views on climate change and energy issues?* Retrieved from <http://www.pewresearch.org/fact-tank/2017/03/22/how-much-does-science-knowledge-influence-peoples-views-on-climate-change-and-energy-issues/>
- Pew Research Center. (2017e). *Highly ideological members of Congress have more Facebook followers than moderates do*. Retrieved from <http://www.pewresearch.org/fact-tank/2017/08/21/highly-ideological-members-of-congress-have-more-facebook-followers-than-moderates-do/>
- Pew Research Center. (2017f). *Partisan conflict and congressional outreach*. Retrieved from <http://www.people-press.org/2017/02/23/partisan-conflict-and-congressional-outreach/>
- Pew Research Center. (2017g). *Republicans skeptical of colleges' impact on U.S., but most see benefits for workforce preparation*. Retrieved from <https://www.pewresearch.org/fact-tank/2017/07/20/republicans-skeptical-of-colleges-impact-on-u-s-but-most-see-benefits-for-workforce-preparation/>
- Pew Research Center. (2018a). *Publics globally want unbiased news coverage, but are divided on whether their news media deliver*. Retrieved from <http://www.pewglobal.org/2018/01/11/publics-globally-want-unbiased-news-coverage-but-are-divided-on-whether-their-news-media-deliver/>
- Pew Research Center. (2018b). *Social media fact sheet*. Retrieved August 31, 2018, from <http://www.pewinternet.org/fact-sheet/social-media/>
- Pew Research Center. (2018c). *Newspapers fact sheet*. Retrieved August 31, 2018, from <http://www.journalism.org/fact-sheet/newspapers/>
- Pew Research Center. (2018d). *Republicans and Democrats agree: They can't agree on basic facts*. Retrieved from <http://www.pewresearch.org/fact-tank/2018/08/23/republicans-and-democrats-agree-they-cant-agree-on-basic-facts/>
- Pew Research Center. (2018e). *Many turn to YouTube for children's content, news, how-to-lessons*. Retrieved from <http://www.pewinternet.org/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/>
- Pigliucci, M., & Boudry, M. (Eds.). (2013). *Philosophy of pseudoscience: Reconsidering the demarcation problem*. The University of Chicago Press. <https://doi.org/10.7208/chicago/9780226051826.001.0001>

- Pinker, S. (2018). *Enlightenment now: The case for reason, science, humanism, and progress*. Penguin Books.
- Pohl, R. F. (Ed.). (2017). *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (Second edition). Routledge. <https://doi.org/10.4324/9781315696935>
- Popan, J. R., Coursey, L., Acosta, J., & Kenworthy, J. (2019). Testing the effects of incivility during internet political discussion on perceptions of rational argument and evaluations of a political outgroup. *Computers in Human Behavior*, 96, 123-132. <https://doi.org/10.1016/j.chb.2019.02.017>
- Porter, T., & Schumann, K. (2018). Intellectual humility and openness to the opposing view. *Self and Identity*, 17(2), 139-162. <https://doi.org/10.1080/15298868.2017.1361861>
- Potegal, M., & Novaco, R. W. (2010). A brief history of anger. In M. Potegal, G. Stemmler, & C. Spielberger (Eds.), *International Handbook of Anger* (pp. 9-24). Springer, New York, NY. [https://doi.org/10.1007/978-0-387-89676-2\\_2](https://doi.org/10.1007/978-0-387-89676-2_2)
- Pritchard, D. (2018). Wittgensteinian hinge epistemology and deep disagreement. *Topoi*. <https://doi.org/10.1007/s11245-018-9612-y>
- Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, 111(3), 781-799. <https://doi.org/10.1037/0033-295X.111.3.781>
- Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28(3), 369-381. <https://doi.org/10.1177/0146167202286008>
- Proust, J., & Fortier, M. (2018). Metacognitive diversity across cultures – An introduction. In Joëlle Proust & Martin Fortier (Eds.), *Metacognitive diversity: An interdisciplinary approach* (chapter 1). Oxford University Press. <https://doi.org/10.1093/oso/9780198789710.003.0001>
- Rabinowitz, M., Latella, L., Stern, C., & Jost, J. T. (2016). Beliefs about childhood vaccination in the United States: Political ideology, false consensus, and the illusion of uniqueness. *PLOS ONE*, 11(7): e0158382. <https://doi.org/10.1371/journal.pone.0158382>
- Ranalli, C. (2018). Deep disagreement and hinge epistemology. *Synthese*, 1-33. <https://doi.org/10.1007/s11229-018-01956-2>
- Ranney, M. A., & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, 8(1), 49-75. <https://doi.org/10.1111/tops.12187>
- Raworth, K. (2017). A doughnut for the Anthropocene: Humanity's compass in the 21<sup>st</sup> century. *The Lancet Planetary Health*, 1(2), e48-e49. [https://doi.org/10.1016/S2542-5196\(17\)30028-1](https://doi.org/10.1016/S2542-5196(17)30028-1)
- Reeves, J. (2019, March). High-profile firing adds to troubles for watchdog group. *The Associated Press*. Retrieved March 23, 2019, from <https://www.apnews.com/11c31b0d4b964a14a069419d8e84e7c6>
- Reinero, D., Wills, J., Brady, W., Mende-Siedlecki, P., Crawford, J., & Van Bavel, J. (2019, February 8). Is the political slant of psychology research related to scientific replicability?. PsyArXiv Preprint. <https://doi.org/10.31234/osf.io/6k3j5>
- Richardson, J. T. E. (2013). Epistemological development in higher education. *Educational Research Review*, 9, 191-206. <https://doi.org/10.1016/j.edurev.2012.10.001>
- Ripple, W. J., Wolf, C., Newsome, T. M., Galetti, M., Alamgir, M., Crist, E., Mahmoud, M. I., Laurance, W. F., & 15,364 scientist signatories from 184 countries. (2017). World scientists' warning to humanity: A second notice. *BioScience*, 67(12), 1026-1028. <https://doi.org/10.1093/biosci/bix125>
- Robichaud, P. (2014). On culpable ignorance and akrasia. *Ethics*, 125(1), 137-151. <https://doi.org/10.1086/677139>
- Robichaud, P. (2016). Is ignorance of climate change culpable? *Science and Engineering Ethics*, 23(5), 1409-1430. <https://doi.org/10.1007/s11948-016-9835-5>
- Robichaud, P., & Wieland, J. W. (Eds.). (2017). *Responsibility – The epistemic condition*. Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.001.0001>

- Rodríguez-Arauz, G., Ramírez-Esparza, N., García-Sierra, A., Ikizer, E. G., Fernández-Gómez, M. J., & José, M. (2018). You go before me, please: Behavioral politeness and interdependent self as markers of simpatía in Latinas. *Cultural Diversity and Ethnic Minority Psychology*, 1-9. Advance online publication. <https://doi.org/10.1037/cdp0000232>
- Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. *Current Biology*, 28(24), 4014-4021.e8. <https://doi.org/10.1016/j.cub.2018.10.053>
- Ronson, J. (2015a, February). How one stupid tweet blew up Justine Sacco's life. *The New York Times Magazine*. Retrieved from <https://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html>
- Ronson, J. (2015b, June). *When online shaming goes too far* [Video]. Retrieved from [https://www.ted.com/talks/jon\\_ronson\\_what\\_happens\\_when\\_online\\_shaming\\_spirals\\_out\\_of\\_control](https://www.ted.com/talks/jon_ronson_what_happens_when_online_shaming_spirals_out_of_control)
- Ronson, J. (2015c). *So you've been publicly shamed* (Updated Paperback Edition). Picador.
- Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives*, 18(1), 295-313. <https://doi.org/10.1111/j.1520-8583.2004.00030.x>
- Rosen, G. (2008). Kleinbart the oblivious and other tales of ignorance and responsibility. *The Journal of Philosophy*, 105(10), 591-610. <https://doi.org/10.5840/jphil20081051023>
- Rosenberg, S. W., & Wolfsfeld, G. (1977). International conflict and the problem of attribution. *Journal of Conflict Resolution*, 21(1), 75-103. <https://doi.org/10.1177/002200277702100105>
- Rosling, H., Rosling, O., & Rönnlund, A. R. (2018). *Factfulness*. Sceptre.
- Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 10, pp. 174-221). New York: Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60357-3](https://doi.org/10.1016/S0065-2601(08)60357-3)
- Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3), 279-301. [https://doi.org/10.1016/0022-1031\(77\)90049-X](https://doi.org/10.1016/0022-1031(77)90049-X)
- Rost, K., Stahel, L., & Frey, B. S. (2016). Digital social norm enforcement: Online firestorms in social media. *PLOS ONE*, 11(6), e0155923. <https://doi.org/10.1371/journal.pone.0155923>
- Rothschild, Z. K., & Keefer, L. A. (2017). A cleansing fire: Moral outrage alleviates guilt and buffers threats to one's moral identity. *Motivation and Emotion*, 41(2), 209-229. <https://doi.org/10.1007/s11031-017-9601-2>
- Royal Society for Public Health (RSPH). (2017). #StatusOfMind – Social media and young people's mental health and wellbeing. Retrieved from <https://www.rsph.org.uk/our-work/campaigns/status-of-mind.html>
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521-562. [https://doi.org/10.1207/s15516709cog2605\\_1](https://doi.org/10.1207/s15516709cog2605_1)
- Rudy-Hiller, F. (2018). The epistemic condition for moral responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2018 Edition. Retrieved from <https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/>
- Russell, B. (1951). *New Hopes for a Changing World*. Simon and Schuster.
- Russell, B. (2008). *Mortals and others – American essays 1931–1935, Volumes I and II*. Routledge Classics. <https://doi.org/10.4324/9780203875339>
- Russell, P., & Deery, O. (Eds.). (2013). *The philosophy of free will: Essential readings from the contemporary debates*. Oxford University Press.
- Rutjens, B. T., Sutton, R. M., & Van der Lee, R. (2018). Not all skepticism is equal: Exploring the ideological antecedents of science acceptance and rejection. *Personality and Social Psychology Bulletin*, 44(3), 384-405. <https://doi.org/10.1177/0146167217741314>

- Saad, L., & Jones, J. M. (2016, March). U.S. concern about global warming at eight-year high. *Gallup*. Retrieved from <https://news.gallup.com/poll/190010/concern-global-warming-eight-year-high.aspx>
- Sagan, C. (1994). *Pale blue dot: A vision of the human future in space*. New York: Random House, Inc.
- Sagan, C. (1996). *The demon-haunted world: Science as a candle in the dark*. Ballantine Books.
- Samuelson, P. L., Jarvinen, M. J., Paulus, T. B., Church, I. M., Hardy, S. A., & Barrett, J. L. (2015). Implicit theories of intellectual virtues and vices: A focus on intellectual humility. *The Journal of Positive Psychology*, 10(5), 389-406. <https://doi.org/10.1080/17439760.2014.967802>
- Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, 114(1), 10-28. <https://doi.org/10.1037/pspa0000102>
- Sandberg, J. (2011). "My emissions make no difference": Climate change and the argument from inconsequentialism. *Environmental Ethics*, 33(3), 229-248. <https://doi.org/10.5840/enviroethics201133326>
- Sartre, J.-P. (1956). *Being and nothingness* (H. E. Barnes, Trans.). New York: Philosophical Library.
- Sawaoka, T., & Monin, B. (2018). The paradox of viral outrage. *Psychological Science*, 29(10), 1665-1678. <https://doi.org/10.1177/0956797618780658>
- Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.
- Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, 26(1-2), 113-125. <https://doi.org/10.1023/A:1003044231033>
- Schroeder, J., Kardas, M., & Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological Science*, 28(12), 1745-1762. <https://doi.org/10.1177/0956797617713798>
- Schuldt, J. P., Konrath, S. H., & Schwarz, N. (2011). "Global warming" or "climate change"? Whether the planet is warming depends on question wording. *Public Opinion Quarterly*, 75(1), 115-124. <https://doi.org/10.1093/poq/nfq073>
- Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S., & Kassam, K. S. (2015). Bias blind spot: Structure, measurement, and consequences. *Management Science*, 61(10), 2468-2486. <https://doi.org/10.1287/mnsc.2014.2096>
- Seli, P., Ralph, B. C. W., Risko, E. F., Schooler, J. W., Schacter, D. L., & Smiler, D. (2017). Intentionality and meta-awareness of mind wandering: Are they one and the same, or distinct dimensions? *Psychonomic Bulletin & Review*, 24(6), 1808-1818. <https://doi.org/10.3758/s13423-017-1249-0>
- Sellars, J. (2002). Marcus Aurelius (121-180 C.E.). In J. Fieser, & B. Dowden (Eds.), *Internet Encyclopedia of Philosophy*. Retrieved March 12, 2019, from <https://www.iep.utm.edu/marcus/>
- Seneca, L. A. (2010). On anger (R. A. Kaster, Trans.). In E. Asmis, S. Bartsch, & M. C. Nussbaum (Eds.), *Anger, Mercy, Revenge* (pp. 1-129). University of Chicago Press.
- Sextus Empiricus. (1996). Outlines of Pyrrhonism (B. Mates, Trans.). In B. Mates (Ed.), *The Skeptic Way*. Oxford University Press.
- Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., Baumeister, R. F., & Vohs, K. D. (2014). Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science*, 25(8), 1563-1570. <https://doi.org/10.1177/0956797614534693>
- Sher, G. (2001). But I could be wrong. *Social Philosophy and Policy*, 18(2), 64-78. <https://doi.org/10.1017/S0265052500002909>
- Sher, G. (2006a). *In praise of blame*. Oxford University Press. <https://doi.org/10.1093/0195187423.001.0001>
- Sher, G. (2006b). Out of control. *Ethics*, 116(2), 285-301. <https://doi.org/10.1086/498464>

- Sher, G. (2009). *Who knew? Responsibility without awareness*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195389197.001.0001>
- Sher, G. (2017). *Me, you, us*. Oxford University Press.
- Sherman, L. W., & Strang, H. (2007). *Restorative justice: The evidence*. The Smith Institute. Retrieved from [http://www.iirp.edu/pdf/RJ\\_full\\_report.pdf](http://www.iirp.edu/pdf/RJ_full_report.pdf)
- Shi, J., Visschers, V. H. M., Siegrist, M., & Arvai, J. (2016). Knowledge as a driver of public perception about climate change reassessed. *Nature Climate Change*, 6, 759-762. <https://doi.org/10.1038/nclimate2997>
- Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics*, 121(3), 602-632. <https://doi.org/10.1086/659003>
- Shotland, R. L., & Heinold, W. D. (1985). Bystander response to arterial bleeding: Helping skills, the decision-making process, and differentiating the helping response. *Journal of Personality and Social Psychology*, 49(2), 347-356. <https://doi.org/10.1037/0022-3514.49.2.347>
- Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspective from the replication crisis. *Annual review of psychology*, 69(1), 487-510. <https://doi.org/10.1146/annurev-psych-122216-011845>
- Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2), 209-215. <https://doi.org/10.1016/j.cognition.2012.04.005>
- Shtulman, A., & Harrington, K. (2015). Tensions between science and intuition across the lifespan. *Topics in Cognitive Science*, 8(1), 118-137. <https://doi.org/10.1111/tops.12174>
- Sidgwick, H. (1981). *The methods of ethics*, 7<sup>th</sup> edition. Indianapolis: Hackett.
- Simpson, B., Willer, R., & Feinberg, M. (2018). Does violent protest backfire? Testing a theory of public reactions to activist violence. *Socius: Sociological Research for a Dynamic World*, 4, 237802311880318. <https://doi.org/10.1177/2378023118803189>
- Singer, P. (2017, August). Why Google was wrong: Did James Damore really deserve to be fired for what he wrote? *Daily News*. Retrieved from <http://www.nydailynews.com/opinion/google-wrong-article-1.3399750>
- Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral intuitions. In J. M. Doris & the Moral Psychology Research Group (Eds.), *The Moral Psychology Handbook* (pp. 246-272). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199582143.003.0008>
- Skaggs, S. L. (2016). Labeling theory. In *Encyclopædia Britannica*. Retrieved from <https://www.britannica.com/topic/labeling-theory>
- Sliwa, P. (2017). On knowing what's right and being responsible for it. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition* (pp. 127-145). Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.003.0007>
- Smith, A. M. (2004). Conflicting attitudes, moral agency, and conceptions of the self. *Philosophical Topics*, 32(1/2), 331-352. <https://doi.org/10.5840/philtopics2004321/27>
- Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics*, 115(2), 236-271. <https://doi.org/10.1086/426957>
- Smith, A. M. (2007). On being responsible and holding responsible. *The Journal of Ethics*, 11(4), 465-484. <https://doi.org/10.1007/s10892-005-7989-5>
- Smith, A. M. (2008). Control, responsibility, and moral assessment. *Philosophical Studies*, 138(3), 367-392. <https://doi.org/10.1007/s11098-006-9048-x>
- Smith, A. M. (2010). Book review – George Sher, *Who knew? Responsibility without awareness*. *Social Theory and Practice*, 36(3), 515-524. <https://doi.org/10.5840/soctheorpract201036327>

- Smith, A. M. (2012). Attributability, answerability, and accountability: In defense of a unified account. *Ethics*, 122(3), 575-589. <https://doi.org/10.1086/664752>
- Smith, D. (2017a, April). Secular Buddhism & secular humanism [Video]. *Secular Buddhist Association*. Retrieved from <https://secularbuddhism.org/secular-buddhism-secular-humanism/>
- Smith, D. (2017b, May). Early Buddhism and secular Buddhism [Video]. *Secular Buddhist Association*. Retrieved from <https://secularbuddhism.org/early-buddhism-and-secular-buddhism/>
- Smith, D. (2017c, June). What did the Buddha teach about views? [Video]. *Secular Buddhist Association*. Retrieved from <https://secularbuddhism.org/what-did-the-buddha-teach-about-views/>
- Smith, D. (2017d, August). What is right speech? [Video]. *Secular Buddhist Association*. Retrieved from <https://secularbuddhism.org/what-is-right-speech/>
- Smith, H. (1983). Culpable ignorance. *Philosophical Review*, 92(4), 543-571. <https://doi.org/10.2307/2184880>
- Smith, I. H., Aquino, K., Koleva, S., & Graham, J. (2014). The moral ties that bind . . . even to out-groups: The interactive effect of moral identity and the binding moral foundations. *Psychological Science*, 25(8), 1554-1562. <https://doi.org/10.1177/0956797614534450>
- Solomon, L., Solomon, H., & Stone, R. (1978). Helping as a function of number of bystanders and ambiguity of emergency. *Personality and Social Psychology Bulletin*, 4(2), 318-321. <https://doi.org/10.1177/014616727800400231>
- Southern Poverty Law Center (SPLC). (2018, June). SPLC statement regarding Maajid Nawaz and the Quilliam Foundation. Retrieved from <https://www.splcenter.org/news/2018/06/18/splc-statement-regarding-maajid-nawaz-and-quilliam-foundation>
- Spence, A., Poortinga, W., & Pidgeon, N. (2012). The psychological distance of climate change. *Risk Analysis*, 32(6), 957-972. <https://doi.org/10.1111/j.1539-6924.2011.01695.x>
- Stanovich, K. E. (2017, July). Were Trump voters irrational? *Quillette*. Retrieved from <https://quillette.com/2017/09/28/trump-voters-irrational/>
- Stanovich, K. E. (2018a). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, 24(4), 423-444. <https://doi.org/10.1080/13546783.2018.1459314>
- Stanovich, K. E. (2018b, July). What is the tribe of the anti-tribalists? *Quillette*. Retrieved from <https://quillette.com/2018/07/17/what-is-the-tribe-of-the-anti-tribalists/>
- Stanovich, K. E., & Toplak, M. E. (2019). The need for intellectual diversity in psychological science: Our own studies of actively open-minded thinking as a case study. *Cognition*, 187, 156-166. <https://doi.org/10.1016/j.cognition.2019.03.006>
- Statman, D. (1993). Introduction. In D. Statman (Ed.), *Moral Luck*. Albany, NY: State University of New York Press.
- Stebay, N. M. (1987). Helping behaviour in rural and urban environments: A meta-analysis. *Psychological Bulletin*, 102(3), 346-356. <https://doi.org/10.1037/0033-2909.102.3.346>
- Stevens, S., & Haidt, J. (2017, August). The Google memo: What does the research say about gender differences? *Heterodox Academy*. Retrieved from <https://heterodoxacademy.org/the-google-memo-what-does-the-research-say-about-gender-differences/>
- Stokes, P. (2012). Philosophy has consequences! Developing metacognition and active learning in the ethics classroom. *Teaching Philosophy*, 35(2), 143-169. <https://doi.org/10.5840/teachphil201235216>
- Strawson, G. (2013). The impossibility of ultimate moral responsibility. In P. Russell & O. Deery (Eds.), *The Philosophy of Free Will: Essential Readings from the Contemporary Debates* (pp. 363-378). Oxford University Press. (Reprinted from *Philosophical Studies*, 75(1-1), pp. 5-24, 1994. <https://doi.org/10.1007/bf00989879> )

- Strawson, P. F. (2013). Freedom and resentment. In P. Russell & O. Deery (Eds.), *The Philosophy of Free Will: Essential Readings from the Contemporary Debates* (pp. 63-83). Oxford University Press. (Reprinted from *Proceedings of the British Academy*, (48)162, pp. 1-25, 1962.)
- Ståhl, T., & van Prooijen, J.-W. (2018). Epistemic rationality: Skepticism toward unfounded beliefs requires sufficient cognitive ability and motivation to be rational. *Personality and Individual Differences*, 122, 155-163. <https://doi.org/10.1016/j.paid.2017.10.026>
- Suhay, E., Bello-Pardo, E., & Maurer, B. (2017). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*, 23(1), 95-115. <https://doi.org/10.1177/1940161217740697>
- Sunstein, C. R. (2018). Growing outrage. *Behavioral Public Policy*, 1-16. <https://doi.org/10.1017/bpp.2018.8>
- Sutherland, S. (1992/2013). *Irrationality: The enemy within*. Pinter & Martin Ltd.
- Taber, C. S., & Lodge, M. (2006). Motivated scepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755-769. <https://doi.org/10.1111/j.1540-5907.2006.00214.x>
- Talbert, M. (2011). Review article – Unwitting behaviour and responsibility. *Journal of Moral Philosophy*, 8(1), 139-152. <https://doi.org/10.1163/174552411x549381>
- Talbert, M. (2013). Unwitting wrongdoers and the role of moral disagreement in blame. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility, Vol. 1* (pp. 225-244). Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199694853.003.0010>
- Talbert, M. (2016). *Moral responsibility*. Polity Press.
- Talbert, M. (2017). Akrasia, awareness, and blameworthiness. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition* (pp. 47-63). Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.003.0002>
- Tauber, S. K., & Dunlosky, J. (2016). A brief history of metamemory research and handbook overview. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory* (pp. 7-22). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199336746.013.13>
- Thoma, N., Pilecki, B., & McKay, D. (2015). Contemporary cognitive behavior therapy: A review of theory, history, and evidence. *Psychodynamic Psychiatry*, 43(3), 423-461. <https://doi.org/10.1521/pdps.2015.43.3.423>
- Thomas, A. K., Lee, M., & Hughes, G. (2016). Introspecting on the elusive: The uncanny state of the feeling of knowing. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory* (pp. 81-94). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199336746.013.16>
- Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 171-196). Oxford University Press. Retrieved from <https://philpapers.org/rec/THODPT>
- Thompson, V. A., Turner, J. P., & Pennycook, G. (2011). Intuition, reason and metacognition. *Cognitive Psychology*, 63(3), 107-140. <https://doi.org/10.1016/j.cogpsych.2011.06.001>
- Thompson, V. A., Evans, St. B. T., & Campbell, J. I. D. (2013). Matching bias on the selection task: It's fast and feels good. *Thinking & Reasoning*, 19(3-4), 431-452. <https://doi.org/10.1080/13546783.2013.820220>
- Thürmer, J. L., & McCrea, S. M. (2018). Beyond motivated reasoning: Hostile reactions to critical comments from the outgroup. *Motivation Science*, 4(4), 333-346. <https://doi.org/10.1037/mot0000097>
- Tognazzini, N. A. (2010). Who knew? Responsibility without awareness [Review]. *Notre Dame Philosophical Reviews*, 2010.01.03. Retrieved from <https://ndpr.nd.edu/news/who-knew-responsibility-without-awareness/>
- Tomperi, T. (2017). Kriittisen ajattelun opettaminen ja filosofia – Pedagogisia perusteita [Philosophy and teaching critical thinking – Pedagogical grounds]. *niin & näin*, 4, 95-112. Retrieved from <https://netn.fi/artikkeli/kriittisen-ajattelun-opettaminen-ja-filosofia-pedagogisia-perusteita>

- Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy & Public Affairs*, 44(3), 197-217. <https://doi.org/10.1111/papa.12075>
- 2veritasium. (2017, March 7). *How should we teach science?* [Video]. Retrieved from <https://www.youtube.com/watch?v=Gstcview6FVM>
- Täuber, S., & van Zomeren, M. (2013). Outrage towards whom? Threats to moral group status impede striving to improve via out-group-directed outrage. *European Journal of Social Psychology*, 43(2), 149-159. <https://doi.org/10.1002/ejsp.1930>
- Täuber, S., van Zomeren, M., & Kutlaca, M. (2015). Should the moral core of climate issues be emphasized or downplayed in public discourse? Three ways to successfully manage the double-edged sword of moral communication. *Climatic Change*, 130(3), 453-464. <https://doi.org/10.1007/s10584-014-1200-6>
- UN (United Nations). (2015). *Transforming our world: The 2030 agenda for sustainable development*. UN General Assembly, New York. Retrieved from <https://sustainabledevelopment.un.org/post2015/transformingourworld>
- UN (United Nations). (2016). *Global sustainable development report 2016*. Department of Economic and Social Affairs, New York. Retrieved from <https://sustainabledevelopment.un.org/globalsdreport/2016>
- Uz, I. (2015). Do cultures clash? *Social Science Information*, 54(1), 78-90. <https://doi.org/10.1177/0539018414554827>
- Vaidis, D. C. (2014). Cognitive dissonance theory. *Oxford Bibliographies Online: Psychology*. <https://doi.org/10.1093/obo/9780199828340-0156>
- Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, 22(3), 213-224. <https://doi.org/10.1016/j.tics.2018.01.004>
- van Bommel, M., van Prooijen, J.-W., Elffers, H., & Van Lange, P. A. M. (2012). Be aware to care: Public self-awareness leads to a reversal of the bystander effect. *Journal of Experimental Social Psychology*, 48(4), 926-930. <https://doi.org/10.1016/j.jesp.2012.02.011>
- van Bommel, M., van Prooijen, J.-W., Elffers, H., & Van Lange, P. A. M. (2016). The lonely bystander: Ostracism leads to less helping in virtual bystander situations. *Social Influence*, 11(3), 141-150. <https://doi.org/10.1080/15534510.2016.1171796>
- Van Dam, N. T., van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Olendzki, A., Meissner, T., Lazar, S. W., Kerr, C. E., Gorchov, J., Fox, K. C. R., Field, B. A., Britton, W. B., Brefczynski-Lewis, J. A., & Meyer, D. E. (2018). Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on Psychological Science*, 13(1), 36-61. <https://doi.org/10.1177/1745691617709589>
- van de Poel, I., Fahlquist, J. N., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics*, 18(1), 49-67. <https://doi.org/10.1007/s11948-011-9276-0>
- van Mill, D. (2018). Freedom of Speech. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2018 Edition. Retrieved from <https://plato.stanford.edu/archives/sum2018/entries/freedom-speech/>
- van Prooijen, J.-W. (2017). Why education predicts decreased belief in conspiracy theories. *Applied Cognitive Psychology*, 31(1), 50-58. <https://doi.org/10.1002/acp.3301>
- van Prooijen, J.-W., & Acker, M. (2015). The influence of control on belief in conspiracy theories: Conceptual and applied extensions. *Applied Cognitive Psychology*, 29(5), 753-761. <https://doi.org/10.1002/acp.3161>
- van Prooijen, J.-W., Krouwel, A. P. M., & Pollet, T. V. (2015). Political extremism predicts belief in conspiracy theories. *Social Psychological and Personality Science*, 6(5), 570-578. <https://doi.org/10.1177/1948550614567356>
- Vargas, M. (2005). The trouble with tracing. *Midwest studies in philosophy*, 29(1), 269-291. <https://doi.org/10.1111/j.1475-4975.2005.00117.x>
- Verhulst, B., Eaves, L. J., & Hatemi, P. K. (2012). Correlation not causation: The relationship between personality traits and political ideologies. *American Journal of Political Science*, 56(1), 34-51. <https://doi.org/10.1111/j.1540-5907.2011.00568.x>

- Veritasium. (2011, March 17). *Khan Academy and the effectiveness of science videos* [Video]. Retrieved from <https://www.youtube.com/watch?v=eVtCO84MDj8>
- Veritasium. (2014, December 1). *This will revolutionize education* [Video]. Retrieved from <https://www.youtube.com/watch?v=GEmuEWjHr5c>
- Vernezze, P. J. (2008). Moderation or the middle way: Two approaches to anger. *Philosophy East and West*, 58(1), 2-16. <https://doi.org/10.1353/pew.2008.0003>
- Vihvelin, K. (2017). Arguments for incompatibilism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2017 Edition. Retrieved from <https://plato.stanford.edu/archives/fall2017/entries/incompatibilism-arguments/>
- Villota, E. J., & Yoo, S. G. (2018). An experiment of influences of Facebook posts in other users. *International Conference of eDemocracy & eGovernment (ICEDEG)*. <https://doi.org/10.1109/ICEDEG.2018.8372319>
- Vincent, E. M. (Ed.). (2018, January). Most popular climate change stories of 2017 reviewed by scientists. *Climate Feedback*. Retrieved from <https://climatefeedback.org/most-popular-climate-change-stories-2017-reviewed-scientists/>
- Visser, P. S., & Mirabile, R. R. (2004). Attitudes in the social context: The impact of social network composition on individual-level attitude strength. *Journal of Personality and Social Psychology*, 87(6), 779-95. <https://doi.org/10.1037/0022-3514.87.6.779>
- Voelkel, J. G., & Feinberg, M. (2017). Morally reframed arguments can affect support for political candidates. *Social Psychological and Personality Science*, 1-8. <https://doi.org/10.1177/1948550617729408>
- Voelpel, S. C., Eckhoff, R. A., & Förster, J. (2008). David against Goliath? Group size and bystander effects in virtual knowledge sharing. *Human Relations*, 61(2), 271-295. <https://doi.org/10.1177/0018726707087787>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151. <https://doi.org/10.1126/science.aap9559>
- Wales, J. (2005, July). *The birth of Wikipedia* [Video]. Retrieved from [https://www.ted.com/talks/jimmy\\_wales\\_on\\_the\\_birth\\_of\\_wikipedia](https://www.ted.com/talks/jimmy_wales_on_the_birth_of_wikipedia)
- Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Cambridge, MA: Harvard University Press.
- Wallace, R. J. (2018). Practical reason. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2018 Edition. Retrieved from <https://plato.stanford.edu/archives/spr2018/entries/practical-reason/>
- Waller, B. N. (2011). *Against moral responsibility*. MIT Press.
- Waller, B. N. (2014). Who knew? Responsibility without awareness, by George Sher [Review]. *Mind*, 123(490), 639-644. <https://doi.org/10.1093/mind/fzu095>
- Walters, D. J., Fernbach, P. M., Fox, C. R., & Sloman, S. A. (2016). Known unknowns: A critical determinant of confidence and calibration. *Management Science*, 63(12), 3999-4446. <https://doi.org/10.1287/mnsc.2016.2580>
- Wang, J., Geng, L., Schultz, P. W., & Zhou, K. (2019). Mindfulness increases the belief in climate change: The mediating role of connectedness with nature. *Environment and Behavior*, 51(1), 3-23. <https://doi.org/10.1177/0013916517738036>
- Washburn, A. N., & Skitka, L. J. (2017). Science denial across the political divide: Liberals and conservatives are similarly motivated to deny attitude-inconsistent science. *Social Psychological and Personality Science*, 9(8), 972-980. <https://doi.org/10.1177/1948550617731500>
- Watson, G. (1975). Free agency. *The Journal of Philosophy*, 72(8), 205-220. <https://doi.org/10.2307/2024703>
- Watson, G. (1996). Two faces of responsibility. *Philosophical topics*, 24(2), 227-248. <https://doi.org/10.5840/philtopics199624222>
- WEF (World Economic Forum). (2018). *The global risks report 2018, 13<sup>th</sup> edition*. World Economic Forum. Retrieved from <https://www.weforum.org/reports/the-global-risks-report-2018>

- Weisberg, D. S., Landrum, A. R., Metz, S. E., & Weisberg, M. (2018). No missing link: Knowledge predicts acceptance of evolution in the United States. *BioScience*, 68(3), 212-222. <https://doi.org/10.1093/biosci/bix161>
- Wesch, M. (2008, June 23). *An anthropological introduction to YouTube* [Video]. Retrieved from <https://archive.org/details/WeschYouTube>
- West, E. J. (2004). Perry's legacy: Models of epistemological development. *Journal of Adult Development*, 11(2), 61-70. <https://doi.org/10.1023/B:JADE.0000024540.12150.69>
- White, F. A., Harvey, L. J., & Abu-Rayya, H. M. (2015). Improving intergroup relations in the internet age: A critical review. *Review of General Psychology*, 19(2), 129-139. <https://doi.org/10.1037/gpr0000036>
- WHO (World Health Organization). (2019). Ten threats to global health in 2019. World Health Organization. Retrieved March 19, 2019, from <https://www.who.int/emergencies/ten-threats-to-global-health-in-2019>
- Wieland, J. W. (2017). Introduction – The epistemic condition. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition* (pp. 1-28). Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.003.0017>
- Wikipedia. (n.d.). In *Wikipedia*. Retrieved October 16, 2018, from <https://en.wikipedia.org/wiki/Wikipedia>
- Williams, B. (1985). *Ethics and the limits of philosophy*. Cambridge, MA: Harvard University Press.
- Williams, H. T. P., McMurray, J. R., Kurz, T., & Lambert, F. H. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32, 126-138. <https://doi.org/10.1016/j.gloenvcha.2015.03.006>
- Wolf, S. (1988). Sanity and the metaphysics of responsibility. In F. D. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 46-62). Cambridge: Cambridge University Press.
- Wolf, S. (1990). *Freedom within reason*. Oxford: Oxford University Press.
- Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, 65, 7-19. <https://doi.org/10.1016/j.jesp.2016.02.005>
- Wong, D. B. (2006). *Natural moralities – A defence of pluralistic relativism*. Oxford University Press. <https://doi.org/10.1093/0195305396.001.0001>
- Wonneberger, A. (2017). Environmentalism – A question of guilt? Testing a model of guilt arousal and effects for environmental campaigns. *Journal of Nonprofit & Public Sector Marketing*, 30(2), 168-186. <https://doi.org/10.1080/10495142.2017.1326873>
- World Bank. (2018). *Poverty and shared prosperity 2018: Piercing together the poverty puzzle*. Washington, DC: World Bank. Retrieved from <https://openknowledge.worldbank.org/handle/10986/30418>
- Wright, R. (2013, November). Why can't we all just get along? The uncertain biological basis of morality. *The Atlantic*. Retrieved from <https://www.theatlantic.com/magazine/archive/2013/11/why-we-fightand-can-we-stop/309525/>
- Wright, R. (2014). Buddhism and modern psychology [Coursera course and video lectures]. *Princeton University, Coursera*. Retrieved May 11, 2019, from <https://www.coursera.org/learn/science-of-meditation>
- Wright, R. (2017). *Why Buddhism is true: The science and philosophy of meditation and enlightenment*. Simon & Schuster.
- Wright, R., & Batchelor, S. (2018, January). *The Wright Show: After Buddhism* [Video and audio file]. Retrieved May 12, 2019, from <https://meaningoflife.tv/videos/39711>
- Wright, R., & Greene, J. (2013, October). *The Wright Show: Robert Wright and Joshua Greene* [Video and audio file]. Retrieved May 11, 2019, from <https://bloggingheads.tv/videos/22741>
- Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological Science*, 18(7), 600-606. <https://doi.org/10.1111/j.1467-9280.2007.01946.x>

- WWF (World Wide Fund for Nature / World Wildlife Fund). (2018). *Living planet report – 2018: Aiming higher*. M. Gooten and R. E. A. Almond (Eds.). WWF, Gland, Switzerland. Retrieved from <https://www.worldwildlife.org/pages/living-planet-report-2018>
- Xu, Z., & Guo, H. (2017). A meta-analysis of effectiveness of guilt on health-related attitudes and intentions. *Health Communication*, 33(5), 519-525. <https://doi.org/10.1080/10410236.2017.1278633>
- Zhou, J. (2016). Boomerangs versus javelins: How polarization constrains communication on climate change. *Environmental Politics*, 25(5), 788-811. <https://doi.org/10.1080/09644016.2016.1166602>
- Zimmerman, M. J. (1986). Negligence and moral responsibility. *Noûs*, 20(2), 199-218. <https://doi.org/10.2307/2215391>
- Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics*, 107(3), 410-426. <https://doi.org/10.1086/233742>
- Zimmerman, M. J. (2008). *Living with uncertainty: The moral significance of ignorance*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511481505>
- Zimmerman, M. J. (2009). Responsibility and awareness. *Philosophical Books*, 50(4), 248-261. <https://doi.org/10.1111/j.1468-0149.2010.00497.x>
- Zimmerman, M. J. (2011). *The immorality of punishment*. Broadview Press.
- Zimmerman, M. J. (2014). Ignorance as a moral excuse [Lecture handout]. Retrieved from <https://philosophy.ceu.edu/events/2014-10-17/ignorance-moral-excuse>
- Zimmerman, M. J. (2017a). Moral responsibility and quality of will. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition* (pp. 219-232). Oxford University Press. <https://doi.org/10.1093/oso/9780198779667.003.0012>
- Zimmerman, M. J. (2017b). Ignorance as a moral excuse. In R. Peels (Ed.), *Perspectives on Ignorance from Moral and Social Philosophy* (pp. 77-94). New York: Routledge.
- Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., & Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLOS ONE*, 12(7), e0181821. Retrieved from <https://doi.org/10.1371/journal.pone.0181821>
- Yang, X., & Dunham, Y. (2019). Minimal but meaningful: Probing the limits of randomly assigned social identities. *Journal of Experimental Child Psychology*, 185, 19-34. <https://doi.org/10.1016/j.jecp.2019.04.013>
- Yagoda, B. (2018, September). The cognitive biases tricking your brain. *The Atlantic*. Retrieved from <https://www.theatlantic.com/magazine/archive/2018/09/cognitive-bias/565775/>

## APPENDICES

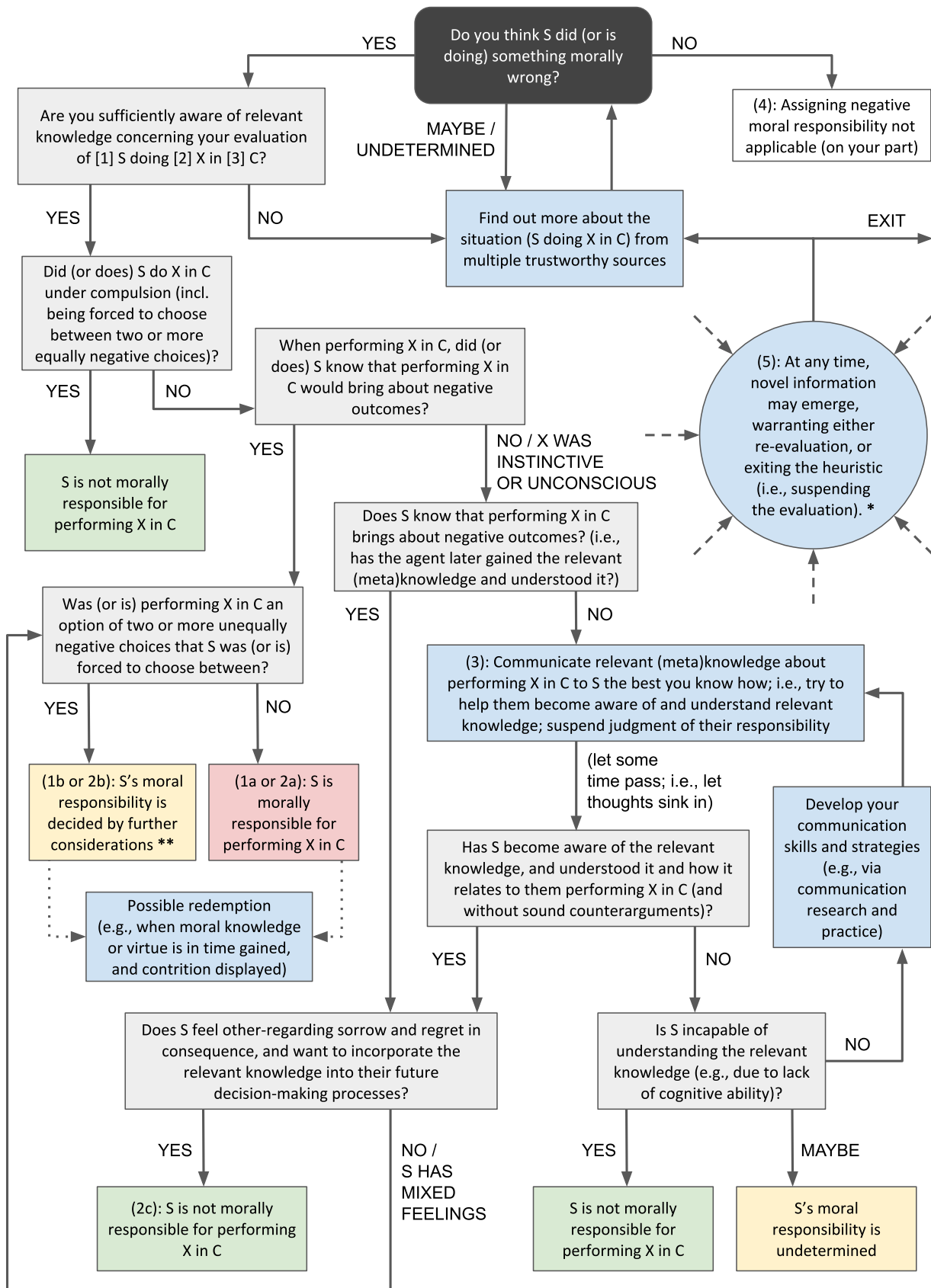
### Appendix 1: Table of Two Intuitive Interpretations of Sher's Example Cases<sup>1</sup>

	<b>Sher:</b> Unaware of:	<b>Sher:</b> Source of cognitive defect:	<b>Sher:</b> Defective judging occurs when:	<b>Metacognitive interpretation:</b> <i>Seemingly</i> unaware of:	<b>Metacognitive interpretation:</b> <i>Possible</i> source of ( <i>apparent</i> ) cognitive defect:	<b>Metacognitive interpretation:</b> <i>(Apparent)</i> defective judging <i>possibly</i> occurs when:
1. <i>Hot Dog</i>	the situation	disruptive thoughts?	distracted?	the situation	unawareness in a relevant preceding situation	lacking metacognitive knowledge
2. <i>On the Rocks</i>	the situation	disruptive thoughts?	distracted?	the situation	unawareness in a relevant preceding situation	lacking metacognitive knowledge
3. <i>Caught off Guard</i>	the situation	disruptive sleepiness?	(getting close to) falling asleep?	the situation	unawareness in a relevant preceding situation	lacking metacognitive knowledge
4. <i>Home for the Holidays</i>	normative demands	distorting emotion (panic)	reacting to panic?	<i>preceding</i> normative demands	unawareness in a relevant preceding situation / distorting emotion (panic via psychological phobia)	lacking metacognitive knowledge or inductive/deductive ability
5. <i>Colicky Baby</i>	normative demands	poor judgment	assessing the facts	( <i>preceding</i> ) normative demands	preceding situation characterised by either not having had access to or not having had ability to access relevant medical knowledge (or not accessing it)	lacking medical knowledge
6. <i>Jackknife</i>	normative demands	poor judgment	assessing the facts	<i>alternative scenarios</i>	inability of precognition, or natural cognitive barriers in making proper split- second decisions	lacking (metacognitive) knowledge or neurological bandwidth needed for a proper evaluation of the quick situation
7. <i>Bad Joke</i>	normative demands	insensitivity to a morally relevant factor	(not) assessing morally relevant factors?	<i>possible</i> normative demands	unawareness of possibly morally relevant factors	lacking knowledge about relevant moral arguments
8. <i>Bad Policy</i>	normative demands	insensitivity to a morally relevant factor	(not) assessing morally relevant factors?	normative demands	preceding situations characterised by not having had access to relevant knowledge about moral psychology	lacking (metacognitive) knowledge about the downsides of emotional empathy
9. <i>Bad Weather</i>	normative demands	poor judgment	formulating moral beliefs	<i>possible</i> normative demands	unawareness of morally relevant factors	lacking knowledge about relevant moral arguments

(section 3.3.2; Sher 2009, 23–29; cf. section 6.1.1.)

<sup>1</sup> The interpretations in the table can be read in the form "agent S was unaware of X due to Y while Z". For example, Sher's interpretation of the scenario *Hot Dog* would thus be: "Alessandra was unaware of Sheba languishing in the car due to disruptive thoughts while being distracted by the school administrators". Similarly, the metacognitive interpretation would be: "Alessandra was unaware of Sheba languishing in the car due to not being aware of the importance of necessary precautions while lacking relevant metacognitive knowledge".

## Appendix 2: Flowchart of the Pragmatic Heuristic for Agent Evaluation



\* Relevant novel information may include, for example, new information about the situation, the target agent or someone else having revealed counterarguments that imply you might be mistaken about what qualifies as relevant knowledge, or simply revealed stress due to the social demands of the heuristic.

\*\* This part of the chart was omitted for the sake of simplicity. For specifics, see sect. 7.1n221 & 7.2.1.