Towards Human-Centered AI that Respects Human Autonomy and Sustainability

KAISA VÄÄNÄNEN, SAARA ALA-LUOPA, ANU LEHTIÖ, THOMAS OLSSON

Tampere University, Faculty of Information Technology and Communication Sciences

kaisa.vaananen@tuni.fi, saara.ala-luopa@tuni.fi, anu.lehtio@tuni.fi, thomas.olsson@tuni.fi

Human-Centered AI (HCAI) has received increasing attention in the past few years. Respecting human values while designing AI applications is one of the core responsibilities of AI developers. In this position paper, we first present the concept of Human-Centered AI (HCAI) and justify its importance in the broader trend of AI development and deployment. We then describe central elements of HCAI from the users' viewpoint, based on our initial literature review of this topic. Finally, we reflect how human autonomy could be supported in AI development and design on two levels: 1) adopting a systematic and holistic design approach which takes into account human values but also broader sustainability goals, and 2) applying more detailed user interface design principles in the AI development while focusing on the HCAI elements.

Keywords and Phrases: Human-Centered AI, autonomy, design, sustainability.

1 INTRODUCTION: WHAT IS HUMAN-CENTERED AI

Although technological advancement is central to the current third wave of AI, there is also increasing attention to the human-centered approach in AI application development. The increasing agency of AI applications challenges developers to think both the ethical viewpoints and the paradigm shift of human-computer interaction where the word interaction could be replaced with integration or collaboration (e.g. [1][2]). As AI technology is becoming more pervasive, we also need to re-think the design elements and the viable methods to bring out the best of both worlds. Human-Centered AI development also calls for strong attention to relevant end-user needs and focus on responsible data use in AI development. [3]

Human-Centered AI (HCAI) has been emphasized as essential to design that helps ensure the acceptance of AI applications. The following descriptions and definitions have been presented:

- Xu [3] argues that HCAI should incorporate three aspects: 1. Ethically aligned design to ensure just, inclusive and fair AI, 2. Technology that fully reflects human intelligence, and 3. Human factors design to ensure that AI solutions are explainable, comprehensible, useful and usable.
- Bond et al. [4] see HCAI as an attempt to bridge HCI research and AI research for the common good of humans. It addresses key human-AI challenges that will enhance the integrity and trust between humans and machines allowing for the appropriate human-machine integration (human-machine teaming).
- HCAI is a perspective on AI and Machine Learning (ML) in which algorithms must be designed with awareness that they are part of a larger system consisting of humans (such as users, operators, clients, and other people in close proximity). HCAI solutions are intelligent systems that are designed with social responsibility in mind, addressing issues such as fairness, accountability, interpretability, and transparency. [5]

General commonalities of HCAI include the objective for ethical AI for 'common good', putting people and their needs at the center of any AI solution, and taking into account the wider sociocultural context. This entails that AI systems should have an understanding of human sociocultural norms as part of their 'theory of mind' about people, and to be capable of producing explanations that non-expert end users can understand [6]. Designing appropriate human-AI collaboration is another central aspect of HCAI [7]. This implies that AI should not be fully autonomous but for many applications, a key is to have "human-in-the-loop". Shneiderman [8] argues that machines and humans are not team mates but ultimately, humans are responsible for the technology they use.

One challenge with AI usage paradigm is that humans are not always aware of being "users" of AI, and hence HCAI must be transparent especially in situations that are critical regarding their consequences. By users we mean people who are actively using a given system, by explicitly giving input or implicitly through their action in certain context(s). Other stakeholders that are affected by a system in some manner without their knowing, action or active involvement should also be considered. For example, a doctor (user) might use a diagnostic system to get consultation in the patient's (stakeholder) medical condition.

2 ELEMENTS OF HUMAN-CENTERED AI

Based on our initial literature review and concept mapping addressing Human-Centered AI, we have found the following elements to be essential in HCAI from the end users' perspective.

Explainability is associated with the notion of explanation as an interface between humans and a decision maker (e.g. AI) that is both an accurate proxy of the decision maker and comprehensible to humans (e.g. [9]). Thus, explainability helps users to understand the AI algorithms and parameters a system uses by, for example, giving a reason for a particular action [3]. Explainability can be a considered as a bridge to avoid unfair and unethical use of algorithmic outputs [10]. Explainability can increase user's understanding by providing information of the essential aspects of the system by utilizing the following elements:

Informativeness. ML models are mainly used in supporting decision making. Still, the problem the AI is trying to solve is not equal to the that of the end-user. Thus, explainable AI should give out information about the (perceived) problem they are tackling, i.e. extracting information about the inner functioning of the AI **Error! Reference source not found.**.

Interactivity can include the ability of an explainable ML model to be interactive with the user (e.g. **Error! Reference source not found.**). This goal is related to fields (e.g. HCI) where the end-users are of great importance and their ability to tweak and interact with the models is seen as ensuring the model's success.

Interpretability is defined as the ability to explain or to provide the meaning in understandable terms to a human (e.g. [10]).

Comprehensibility refers to the ability of a learning algorithm to represent what it has learned/ what it knows in a way that humans can understand. Given that this is difficult to quantify, comprehensibility is normally tied to the evaluation of the model complexity (e.g. [9])

Beyond explainability and its attributed listed above, there are further HCAI aspects that developers need to aim at. ***Transparency*** can be seen as an antecedent of explainability. Transparency provides means to access the datasets and workflows inside an AI system to end users [5], increasing a sense of trust. Arrieta et al. [10] state that transparency is the opposite of 'black-box-ness'. Transparency is also closely linked to ***fairness***, i.e. the requirement that all users are treated equally and without prejudice. ***Trustworthiness*** can be considered as the confidence of whether AI will act as intended when facing a given problem. Trustworthiness is also not easy to quantify nor is the relationship between explainability and trustworthiness reciprocal. [10] ***Privacy awareness*** is an AI's ability to assess privacy. AI models can have complex representations of their learned patterns and not being able to understand what the model has captured and stored can entail a privacy breach (e.g. [12]).

3 DESIGNING FOR HUMAN AUTONOMY IN AI APPLICATIONS

While the concept of autonomy is central to AI applications – in the way they function (partly) independently of the user input or without people's explicit knowledge – there has been less attention to the autonomy of the people using or affected by AI. Autonomy is a central human need [13] and sense of autonomy is necessary for people's wellbeing. It can be defined as ***"an individual's capacity for self-determination or self-governance"*** (Internet Encyclopedia of Philosophy, https://iep.utm.edu/autonomy/). AI has increasing agency which implies some level of autonomy. This can be considered an advantage of intelligent technology solutions, but it may also lead to violation of human autonomy. While these two perspectives to autonomy are important, we would like to bring into the equation the pressing need to consider various sustainability goals (such as UN's 17 goals, https://www.un.org/sustainabledevelopment/sustainable-development-goals/) in all development efforts. While human autonomy is important to consider over system autonomy, the "autonomy" of our globe needs to be put into focus as well.

In the following we discuss how human autonomy can be supported on two levels: (3.1) in broader design approaches used to aim at human-centeredness and sustainability in AI development and (3.2) with specific design considerations that can support central HCAI elements, in order to achieve AI applications that respect human autonomy.

3.1 Design Approaches for HCAI

Respecting human autonomy in the overall design approach requires that users and other stakeholders are put in control over what is developed for them. People should have a say how systems affecting their lives are designed. The potentially highly disruptive and far-reaching influences of AI systems call for new sensitivities and perspectives in the practice of HCI design. Environmental and social sustainability should be in the core of all developers' responsibility.

Human-Centered Design (HCD) [14] is a viable starting point to HCAI. HCD involves different stakeholders actively in the design process, and utilizes an iterative approach to understand user needs, and evaluate the user experience and effects of the designs. For HCAI, the basic methods of HCD may need to be adapted to account for the evolving nature of AI, for example by extensive data-based testing to catch potential biases, or to focus on the explainability issues described above. ***Value-sensitive design*** [15] builds upon HCD but goes deeper into understanding people's values and using their nuances to guide the design. In HCAI, human autonomy as a value and the associated need to maintain control in human-AI interactions may imply that "human-in-the-loop" becomes a central goal in design.

Sustainable design [16], ***speculative design*** [17], ***socially responsible design*** [18] and ***planet-centered design*** (https://planetcentricdesign.com) can replace or complement the conventional human-centric design approach, in order to achieve solutions that are sustainable. We argue that future AI designers need competences that allow them to more holistically reflect on the values and assumptions underlying proposed solutions, to elicit discussion on designers' responsibility with the help of speculative artefacts, to conceive artefacts that are capable of communicating ideas with psychological and ideological weight, and to envision alternative socio-technical futures where AI technology can impact different behavioral, cultural and societal dynamics. In other words, rather than tools for crafting, such approaches provide tools for thinking in new ways and being mindful of why we design things.

3.2 User Interface Design Principles for Human Autonomy

The very central target of human-centered design is to design applications with purposeful functionality, leading to effectiveness of system use [14]. Since AI functions (at least partly) autonomously and hence takes actions on behalf of the stakeholder, ***user control*** (real and perceived) is a very central aspects to consider in HCAI. Further

central considerations are ***appropriate task allocation*** to support purposeful human-AI collaboration and ***respecting privacy*** of various stakeholders.

Other, more detailed user interface (UI) design principles can be extracted from various design heuristics and guidelines (e.g. by Jacob Nielsen and Don Norman) such as ***system status visibility, informative feedback, speaking the user's language, error prevention and recovery, user guidance, recognition over recall, reversal of actions,*** and ***providing task closure***. Some of the well-known UI design principles may be harder to follow in AI applications, such as ***consistency*** and ***simplicity***, especially if at the same time trying to achieve adequate explainability.

How these design considerations should be addressed depends on the stakeholders' (target groups') needs and skills, as well as on the application domain. Regarding sustainability, such low-level UI design choices may in their part advance social sustainability by offering usable AI solutions that can be used by versatile user groups, including the non-experts.

ACKNOWLEDGMENTS

We are grateful for Tampere Region Central Fund, European Union Regional Fund, Business Tampere, and Tampere University for funding this research through KITE project (Human-Centered AI Solutions for the Smart City).

REFERENCES

[1] Farooq, U., & Grudin, J. (2016). Human-computer integration. ***interactions***, ***23***(6), 26-32.

[2] Johnson, M., Bradshaw, J. M., & Feltovich, P. J. (2018). Tomorrow's human–machine design tools: From levels of automation to interdependencies. ***Journal of Cognitive Engineering and Decision Making***, ***12***(1), 77-82.

[3] Xu, W. (2019). Toward human-centered AI: a perspective from human-computer interaction. interactions, 26(4), 42-46.

[4] Bond, R. R., Mulvenna, M., Finlay, D., Wong, A., Koene, A., Brisk, R., ... & Adel, T. (2019). Human Centered Artificial Intelligence: Weaving UX into Algorithmic Decision Making. In RoCHI 2019: International Conference on Human-Computer Interaction.

[5] Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. ***Human Behavior and Emerging Technologies***, 1(1), 33-36.

[6] Ford, K. M., Hayes, P. J., Glymour, C., & Allen, J. (2015). Cognitive orthoses: toward human-centered AI. ***AI Magazine***, 36(4), 5-8.

[7] Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. ***In EA of the CHI 2020***, ACM.

[8] Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. ***International Journal of Human–Computer Interaction***, 36(6), 495-504.

[9] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ***ACM computing surveys (CSUR)***, ***51***(5), 1-42.

[10] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. ***Information Fusion***, 58, 82-115.

[11] Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017, February). Explainable agency for intelligent autonomous systems. In ***Twenty-Ninth IAAI Conference***.

[12] Castelvecchi, D. (2016). Can we open the black box of AI?. ***Nature News***, ***538***(7623), 20.

[13] Ryan, R. M. & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. ***American Psychologist***. 55: 68–78. doi:10.1037/0003-066X.55.1.68

[14] ISO 9241-210:2010. Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.

[15] Friedman, B., Hendry, D.G. & Alan Borning (2017). A Survey of Value Sensitive Design Methods. ***Foundations and Trends in Human–Computer Interaction***: Vol. 11: No. 2, pp 63-125. http://dx.doi.org/10.1561/1100000015

[16] McLennan, J. F. (2004), The Philosophy of Sustainable Design.

[17] Dunne, A. & Raby, F. 2013. Speculative Everything: Design, Fiction, and Social Dreaming. The MIT Press.

[18] Melles, G., de Vere, I. & Misic, V. (2011) Socially responsible design: thinking beyond the triple bottom line to socially responsive and sustainable product design. ***CoDesign***, 7:3-4, 143-154, DOI: 10.1080/15710882.2011.630473