





# 9

## Toward Designing Ethically Acceptable AI Security Systems Through Agent Modeling

Jaana Hallamaa , Tomi Janhunen ,  
Jyrki Nummenmaa , Timo Nummenmaa ,  
Pertti Saariluoma, and Elizaveta Zimina

### Introduction

Security is a crucial concern in public places such as shopping malls. People need to feel safe and businesses should run smoothly; hence, the security measures should be sufficient but not too exaggerated. Overall, public places and shopping mall security present a complicated topic, as practically everything starting from the building design is relevant. Harmful events such as violent attacks or overreaction from guards will reduce interest in visiting a shopping mall.

---

J. Hallamaa  
University of Helsinki, Helsinki, Finland  
e-mail: [jaana.hallamaa@helsinki.fi](mailto:jaana.hallamaa@helsinki.fi)

T. Janhunen  
Tampere, Finland  
e-mail: [tomi.janhunen@tuni.fi](mailto:tomi.janhunen@tuni.fi)

AI has many components that make it useful in monitoring security in a public place. Conducting event analysis from videos, various sensor data, and voice data is a challenging task; hence, the utilization of AI becomes inevitable. Building AI systems requires considerable amounts of human and computational resources. Therefore, the suitability of such AI systems should be studied in advance.

For this purpose, in this chapter, we propose the employment of modeling of relevant actors and shed light on the ethical concerns surrounding them as a multi-agent system (MAS). MASes serve as fundamental models for AI systems and their operating environments, offering flexible means for their definition, analysis, and implementation through agent languages. When understanding agents' behavior, beliefs, desires, and intentions (BDI) are central concepts that have been widely applied in literature. In this chapter, the moral dimensions of BDI agents are considered. We approach them from the perspective of interaction and presuppose cooperation between agents that is based on social intentionality, thus initiating a framework for the socio-ethical modeling of agency. The framework utilizes three modes of social interaction that can be attributed to the intentions of participating agents. Throughout the chapter, social phenomena and scenarios arising within the context of a shopping mall are employed to drive discussion and analysis. In addition to theoretical considerations, the premises for practical implementations are defined in a GAMA model. Simulations and visualizations created using the proof-of-concept implementation serve to illustrate and to communicate the model for stakeholders.

MASes provide abstract models of AI systems in action, ranging from complex societies of collaborative and/or competitive agents to simple single-agent problem-solving scenarios (see Woolridge, 2009, for a

---

J. Nummenmaa (✉) • T. Nummenmaa • E. Zimina

Tampere University, Tampere, Finland

e-mail: [jyrki.nummenmaa@tuni.fi](mailto:jyrki.nummenmaa@tuni.fi); [timo.nummenmaa@tuni.fi](mailto:timo.nummenmaa@tuni.fi)

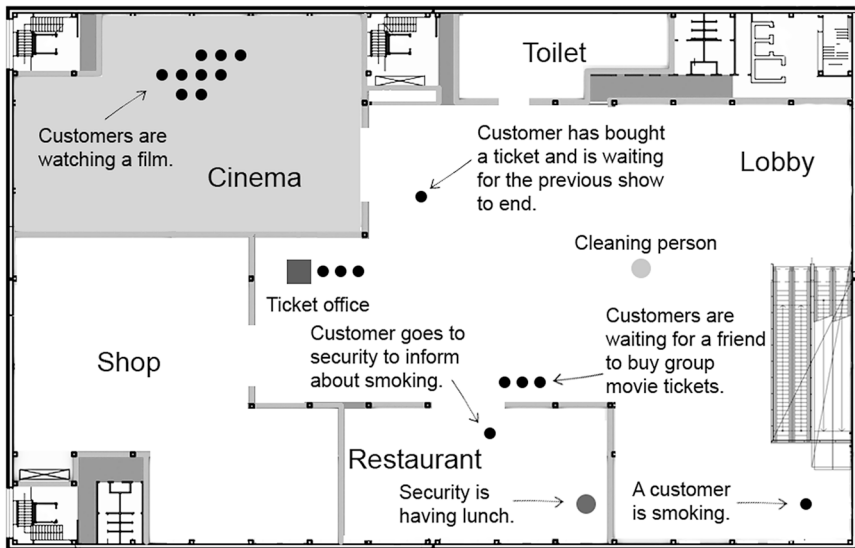
P. Saariluoma

University of Jyväskylä, Jyväskylä, Finland

e-mail: [ps@jyu.fi](mailto:ps@jyu.fi)

comprehensive introduction). Regardless of the application, individual agents in such systems perform actions in response to the perceptions they gather from their environments. An *ideal* and *rational* agent is expected to achieve its goals and maximize its *expected utility* in the long run (Russell & Norvig, 2020). To analyze moral aspects, agents should possess functionalities beyond perceiving and acting. A widely accepted approach defines *beliefs* (B), *desires* (D), and *intentions* (I) as fundamental elements of the behavioral description of agents (Rao & Georgeff, 1991). The BDI architecture serves as a solid foundation for addressing ethically relevant settings within MASes.

Our empirical case concerns a shopping mall illustrated in Fig. 9.1 and, more specifically, the security, monitoring, and maintenance activities of the mall. The example includes only an abstraction of a particular section of a real mall, with only a fraction of its services and activities. Nevertheless, it serves us as a challenging environment that is easy to understand in general but presents endless possibilities for refinement from the modeling perspective. In this light, our series of examples will



**Fig. 9.1** Illustration of the simulation environment: a shopping mall. *Source:* Authors (2023)

specifically concentrate on modeling the activities of customers and staff members (security guards in particular). In general, agents may participate in activities individually or by collaborating with others. Their actions may give rise to ethical concerns, which constitute the particular focus of our research. Some actions and activities in the mall involve groups of agents committed to joint goals, such as going to the movies together. The ways in which groups form and organize themselves vary, increasing the complexity of the process and the need for social interaction. While group formation is a complex process in itself, it is not the main focus of this chapter. Instead, we take into account the roles played by groups of agents (if present).

In this chapter, we address the ethically relevant or moral aspects of MASes. Our overall goal is to find suitable primitives for the formalization of ethical principles and, in this manner, establish the premises for ethical modeling in multi-agent contexts. Ideally, ethical principles can be separated from operational details, and once formalized, they can be employed to analyze and answer ethically relevant questions. Our interdisciplinary approach emphasizes philosophical aspects, aiming to gain a fundamental understanding of ethically meaningful primitives from the analysis of multi-agent scenarios in the mall domain.

Rather than concentrating on the representation of norms (see, e.g., Broersen et al., 2001; Neumann 2010), we take the *modes of social intentionality* (Tuomela, 2007) as a starting point for our analysis, thereby adopting a *socio-ethical* approach to modeling MASes. Our long-term goal is to facilitate the implementation of MASes and their simulation as well as promote the development of agent (specification) languages. However, we do not introduce new languages in this preliminary study and instead utilize an existing one, namely, GAMA (Taillandier et al., 2019), in our illustrations and proof-of-concept implementations.

To summarize, this chapter initiates a socio-ethical viewpoint and approach to modeling BDI agency. Its main contributions include the following:

1. establishing the framework of *modes of social intentionality* for the analysis of BDI agency;
2. analyzing the grounds for moral action on the basis of the modes of action of the agents involved;

3. applying the framework in the socio-ethical modeling of an open-ended application domain (the mall domain); and
4. addressing the limitations of traditional BDI models in the formalization of ethical principles.

## Multi-agent Systems

A MAS constitutes an ecosystem of computing entities, namely, *agents*, each of which solves some sub-problem as part of a larger collective endeavor. The agents in a MAS form a network by virtue of sharing knowledge and communicating with each other. Other capacities, say the ability to follow if-then rules and core behaviors such as mobility, interaction, adaptation, and learning, have been listed as their characteristics (see, e.g., Balaji & Srinivasan, 2010; Rocha et al., 2017).

## Agents and Environments

An *agent A* is an entity whose *state* consists of precisely defined mental components such as beliefs, capabilities, choices, and commitments that roughly correspond to their common-sense counterparts in humans (Shoham, 1993). Additionally, *values* that are more concrete may also be relevant when characterizing states. In the field of computer science, it is a common practice to formalize the states of agents by introducing *state variables* whose values range over particular domains of interest.

The properties of the environment of a MAS are essential when it comes to designing a MAS in the first place, and they also determine how difficult it is for the MAS to achieve its goals. Environments can be roughly classified on the basis of their central characteristics, allowing us to define ranges such as *static* versus *dynamic*, or *fully observable* versus *partially observable* (Russell & Norvig, 2020). In simple MASes, it is also possible to view the environment as one agent *hosting* others (cf. typical *master-slave* architectures).

## Actions

Agents in a MAS interact with each other and their environment by performing *actions*. The actions serve two primary purposes: either *observing* or *changing* the state of the MAS, which includes the states of the individual agents as well as that of the environment. In addition to actions performed by agents, *events* occurring unexpectedly in the environment may also affect the state of the MAS. The (effects of) actions and events can be defined in different ways, for example, by assigning new values to state variables on the basis of old ones.

In logic-oriented formalisms, such as STRIPS (Fikes & Nilsson, 1971), and the so-called *action languages* (Gelfond & Lifschitz, 1998), the states of a system can be described using *state predicates*, also known as *fluents*, whose truth values may change over time. The same applies to *actions*: A particular action can be performed if its *preconditions* are met. As the result of executing an action, certain fluents may receive truth values, thereby establishing the *postconditions* of the action. We describe changes such as these either as *additions* or *deletions* of predicates, which are sufficient to cover four possible cases for each fluent, namely, whether it stays/becomes true/false.

### Example 9.1

Consider a customer  $C$  entering the mall. Let  $e$  and  $l$  be the names denoting the entrance and the lobby of the mall, respectively. Furthermore, let predicates  $next/2$  and  $in/2$  describe whether a customer is next to something or in a particular space.

**Action**  $enter(C)$ :

- *Precondition(s)*:  $next(C, e)$ .
- *Addition(s)*:  $in(C, l)$ .
- *Deletions(s)*:  $next(C, e)$ .

Consider a particular customer  $c_1$  at the entrance, that is,  $next(c_1, e)$  is true, thus enabling the action  $enter(c_1)$ . When executed,  $next(c_1, e)$  is falsified while  $in(c_1, l)$  becomes true.

## Group Actions

In Example 9.1, the action involves a single agent. As a result, the state of the agent changes as reflected by the modified truth values of the fluent involved. In multi-agent scenarios, we consider group actions engaging several agents.

### Example 9.2

*Continuing our examples, consider the act of one customer  $C_1$  approaching another ( $C_2$ ) in the same space  $S$ .*

**Action**  $\text{approach}(C_1, C_2)$ :

- *Precondition(s)*:  $\text{in}(C_1, S), \text{in}(C_2, S)$ .
- *Additions(s)*:  $\text{next}(C_1, C_2), \text{next}(C_2, C_1)$ .

*As a result, both customers remain in the space  $S$ , but they appear next to each other afterward as encoded with the fluent  $\text{next}/2$ .*

As a result of a group action, the states of all agents involved may be updated. The action in Example 9.2 is asymmetric by nature, and the latter agent is merely treated as an object. The other agent might react by escaping from the situation by performing a counteraction  $\text{escape}(C_2, C_1)$ , thus falsifying the fluents  $\text{next}(C_1, C_2)$  and  $\text{next}(C_2, C_1)$ . These conditions are the natural pre- and postconditions for yet another group action:  $\text{shake-hand}(C_1, C_2)$ .

## BDI Models Formalized

Fluents describing a MAS essentially express the components of its state that are relevant for modeling. As usual, the meaning of such predicates can be decided on a case-by-case basis. For instance, in our shopping mall domain, if  $\text{in}(C, S)$  is true for a particular customer  $C$  and a space  $S$ , then

$C$  is in  $S$ . Ethical aspects, however, cannot be directly formalized using state predicates, since the mental states of agents matter as well. To this end, one prevailing approach captures the *beliefs*, *desires*, and *intentions* of agents as meta-level concepts. In the sequel, we follow Labrou and Finin (1994) and formalize these concepts in terms of *modal operators*  $B_A$ ,  $D_A$ , and  $I_A$  associated with an agent  $A$ . For now, we restrict the application of these operators only to fluents or their negations, hence forbidding nesting. This is primarily to mitigate computational complexity and facilitate implementation.

### Example 9.3

*Consider a customer  $C$  who wants to see a movie  $M$  in a particular theatre  $t$  of the mall.*

**Action** buy-ticket( $C, M$ ):

- *Precondition(s)*:  $I_C(\text{watch}(C, M))$ ,  $\text{in}(C, t)$ .
- *Additions(s)*:  $\text{has-ticket}(C, M, t)$ .

*Ticket possession is one of the natural preconditions for seeing a movie.*

**Action** watch( $C, M$ ):

- *Precondition(s)*:  $\text{in}(C, t)$ ,  $\text{has-ticket}(C, M, t)$ ,  $I_C(\text{watch}(C, M))$ ,  $\text{showtime}(C, M, t)$ .
- *Deletions(s)*:  $\text{has-ticket}(C, M, t)$ ,  $I_C(\text{watch}(C, M))$ .

*In the above, the mental state of the customer is updated accordingly, that is, the intention of seeing the movie is falsified.*

The management of desires is an independent aspect: Persistent desires can be maintained indefinitely, while those that are more one-time by nature can be abandoned by falsifying  $D_C(D)$ .

## Philosophy of Action in BDI Models

Two conditions determine whether an agent  $A$  succeeds in performing an action:  $A$  must succeed in performing the planned act and, second, the effects of the act must fulfill or further the premeditated goal. The concept of human intentionality implies a *conational* and an *epistemic* attitude in  $A$ : The goal of the intended action is something the agent desires, wishes, or wants to achieve or make real, and the agent knows, believes, or hopes that the intended action is a means to realize it. The BDI model that conceptualizes agency in terms of beliefs, desires, and intentions fundamentally aligns with these conditions. Thus, by providing a MAS with a set of agents that fulfill the conditions of the BDI model, we can establish a connection between the MAS and the philosophical concept of human action (Adam & Gaudou, 2015).

BDI agents have been analyzed in relation to their usability in different types of social simulations (Adam & Gaudou, 2015). The advantage of utilizing a BDI model is that it supports a large variety of agent architectures, such as a particle- or a rule-based architecture, a neural network, or a cognitive architecture.

By designing different types of cognitive architectures within the BDI framework, it is possible to simulate reasoning, norm-based behavior, and decision-making processes as well as study the effects of interaction between agents. Agents modeled according to BDI are more akin to real human beings and are better at mimicking their behavior than agents based on the psychological concepts of cognitive science. The BDI model concentrates on the conscious and the observable level of agent behavior instead of focusing on the—often—unconscious psychological states of the agent. The model provides a common-sense understanding of how desires, currently held information, and communication with others affect behavior (Adam & Gaudou, 2015).

## Modes of Social Action

The conceptual tools for addressing human MASes stem from the philosophy of sociality. We can distinguish between different types of cooperation depending on  $A$ 's intentionality and attitude toward other agents

in the group. Following Tuomela (2007), we can distinguish the following three modes of multi-agent cooperation on the basis of social intentionality: *pure I-mode*, *progroup I-mode*, and *we-mode* social intentionality.

The differences between the three modes of intentionality become clear when we highlight the relationship between  $A$  and the other agents involved in cooperation. The group of other agents may be described as a *surrounding*, an *instrument*, or an *end in itself* for  $A$ , depending on  $A$ 's mode of intentionality. We sum up the specific features of the three modes in what follows:

In *pure I-mode action*,  $A$  acts within the group that provides a *surrounding* for the actions of its members.

#### **Example 9.4**

*By entering the shopping mall, customer  $C_1$  becomes a member of a pure I-mode group, see Example 9.1.*

Group membership in pure I-mode action is based on each agent's (some) individual intention that happens to be similar to (at least one of) the current individual intentions of the other agents. Customers at a shopping mall constitute a group in terms of their individual intentions to be at the shopping mall at a given moment. What other agents do affects  $A$ 's conditions of actions, and each may enter or leave the mall for their individual reasons and as they wish. Agent  $A$  may sabotage pure I-mode cooperation by preventing others from participating in the common activity, as would happen if  $A$  started to pester other customers at the mall.

Cooperation in *progroup I-mode* presupposes that all members of the group commit themselves, first, to the same goal and, second, to each other and each other's part in the cooperation. Acting in the group offers *instrumental value* to its members, as each of them can achieve their individual goal—which is common for all members—better by acting in the group than trying to realize it alone. Goal achievement may involve *division of labor* by *delegating* and *dividing* tasks among subgroups or individual members. AI identifying these *roles* enables the group as a whole to perform *concurrent* actions.

Acting in a group now presupposes that  $A$  does not act in a *counter-productive* manner, for example, by hindering others from achieving the common goal or by sabotaging the initiatives of other group members, as would happen if say customer  $C_2$  turns down customer  $C_1$ 's offer to shake hands (see the text following Example 9.2.) Customer  $C_2$ 's refusal to shake hands would also prevent  $C_1$  from shaking hands, thereby ending an instance of progroup I-mode action.

The group may adopt several goals that must be set in a priority order for their effective realization. Then  $A$  may have to *compete* with other group members as to whose current interests will be given most weightage. As a member of this type of a group,  $A$  may have to work toward reaching aims other than  $A$ 's individual goals. Cooperating in the group is a price that  $A$  pays for the instrumental value of the group as a *means* to ensure *reciprocity* between the members of the group. By agreeing to further the group's aims,  $A$  has a better chance of getting the group to work toward an aim closer to  $A$ 's interests. Such compromises and trade-offs are signs of valuing *fairness* in cooperation.

### Example 9.5

*A group of friends has agreed to meet at the shopping mall to get movie tickets at a group discount price for a film they all wish to see. The group members share  $A$ 's aim to see the film and going together has instrumental value: The ticket is cheaper. The group may divide tasks as to who buys the tickets, who takes care of snacks, and who reserves a table at a restaurant after the movie.*

*We-mode intentionality* involves  $A$  having two intended goals: first, to take part in realizing the aim that justifies the existence of the group, that is, the group goal, and second, to do  $A$ 's part in keeping the group together by enabling its members to act as a group. The group is now also an *end in itself* for  $A$ , not just an instrument to realize a goal. This type of cooperation requires strong and often long-term commitment, as the group members must commit themselves both to the shared aim and to the group as a totality, as well as to the members of the group as parts of the whole.

**Example 9.6**

*A team of security guards often functions in we-mode. They share the interest of each security guard  $S$  doing their own part in the job during their shift as well as help their coworkers execute their part according to the principle: one for all and all for one. The team is not just an instrument for each  $S$  to execute their duties but has value in itself for its members.*

Distinguishing between the different modes of social action helps explicate the significance that other group members' desires, intentions, and attitudes hold for  $A$  and determine  $A$ 's possibilities of realizing the desired goals. Action based on cooperation becomes impossible if those who form a group cannot *trust* each other. Although there are different ways to formalize trust, the common feature among all trust models is that they are computational methods used for calculating  $A$ 's trust in a trustee  $T$ , (see Koster et al., 2013).

The three modes of social action constitute a theoretical model for analyzing different types of cooperation in terms of the bonds holding the agents together as members of a group and their commitments to each other, the group as a whole, and the cooperation. In real life, people mostly act according to unspoken but internalized social conventions and practices that direct human behavior. People are raised and socialized to follow culturally determined norms, and they comply with them even in a crowd where encounters between individuals are random. The three modes of social action all deal with positive instances of cooperation, presupposing that partaking in social action fulfills some individual interest of each agent. Actions based on progroup I-mode and we-mode represent cooperation in which the group members commit themselves to the common task, aiming for the desired outcome and showing their commitment to each other. We would require different types of conceptual tools for addressing offensive and negative action as the three modes of social intentionality discussed above do not cover such situations.

As the modes each provide a distinct way to display different degrees of social intentionality, we have to integrate them into our modeling process. First, we have to choose between implicit and explicit modeling. To

discuss the properties of these two options, we present a simple case where the mode changes.

### Example 9.7

*It is almost noon and a security guard decides to take a lunch break. The guard, currently acting in we-mode, switches to pure I-mode for the duration of the lunch break to take a break from work activities. While in pure I-mode, the guard is not interested in the tasks of the group of guards; however, that task is still dormant in the background. If a relevant event (e.g., emergency) were to take place during the lunch break, the guard might switch back to we-mode, joining the operations of the group of guards.*

We can attempt to model this example implicitly, that is, by deciding that each intention of the security guard is related to a specific mode. Therefore, the change of intention would also signify a change of mode, and the actions that change the state of the MAS would be written in a manner that takes the modes into account. The modes would be present implicitly in the definition of actions but not explicitly defined. For example, each action that takes the security guard closer to eating would not include the execution of work tasks for sure. There is an issue here, however. We may wish an agent to reach the same target condition of an intention in different ways by acting in various modes. The solution is to explicitly record the mode of action when committing to an intention.

### Example 9.8

*Security guard  $S$  commits to having lunch in some restaurant  $R$  of the mall in pure I-mode ( $im$ ):*

**Action** `commit_to_lunch( $S, R$ ):`

- *Precondition(s):*  $\mathcal{D}_S(\text{lunch}(R))$ .
- *Addition(s):*  $\mathcal{I}_S(\text{lunch}(R))$ ,  $\text{mode}(S, \text{lunch}(R), im)$ .

## Moral Action

In the context of human agency, an agent  $A$  must be *morally responsible*: People are blamed and praised for what they do. Moral blame implies an obligation to repair the harm caused and to ask those who have been harmed for forgiveness. The society prosecutes and punishes those who engage in such harmful deeds according to its legislation. Morally praiseworthy actions are favorable for others and deserve positive acknowledgment. As such, responsibility is too strict a condition for non-human agents (Hallamaa & Kalliokoski, 2020).

From the point of view of moral consideration, based on von Wright (1968), the goal of any action represents a *value*, which is something an agent regards as good, beneficial, or favorable in relation to its (present) interests. In general, goals that are beneficial for other agents, too, are *morally good*. If realizing the goal does not affect other agents' well-being, it is *morally neutral*. Morally good and neutral goals are *morally permissible*. Goals that directly harm other agents are *morally bad* and can even be defined as *morally evil*, if the harm is intended by (part of)  $A$ 's action. It is morally *forbidden* for  $A$  to set such goals and to try to reach them through actions.

Actions, too, can be divided into three categories depending on their *moral permissibility*. In general, acts that are good, beneficial, or favorable in terms of their consequences to other agents are morally good, acts that do not affect the well-being of others are morally neutral, and acts that harm other agents are morally bad or evil. The morally good and neutral acts belong to the category of permissible acts, whereas the morally bad acts are classified as forbidden acts. Acts often yield different outcomes for different parties depending on their position in the situation. The same act can thus be both favorable and unfavorable. If the act itself is not forbidden, assessing its moral value often includes weighing the outcomes for those involved.

### Example 9.9

*The customers are free to choose their goals from the set of actions that do not harm or hinder the functioning of the shopping mall, which, normally, consist*

*of morally neutral acts such as making purchases, enjoying a meal, and resting one's feet on a bench. Likewise, the customers should not (try to) do anything that would inhibit or hinder other customers and staff from setting their own goals and performing acts that are appropriate instances of behavior in the shopping mall context.*

In exceptional cases, morally forbidden acts may be permissible if *A*'s aim is to preserve something of (great) value, and the likely outcome of the harmful action is (expected to be) more positive than the anticipated outcome of *A* not performing the action.

### **Example 9.10**

*A security guard  $S$  may use physical force to hinder customer  $C_1$  from punching  $C_2$ .*

Some of the permissible acts are *required* or *compulsory*, and *A* has a moral *obligation* to perform them in a certain situation or context. A conceptual connection exists between what is permissible and compulsory in the following manner: All compulsory acts are permissible, and none of the impermissible acts are compulsory.

### **Example 9.11**

*The customers must finish their purchases and leave the shopping mall when the closing time is approaching. The guards have an obligation based on their duties to ensure that the customers leave the premises. The same applies to emergency situations: The sounding of the fire alarm indicates that the customers must leave the mall immediately, disregarding what they are doing, and the guards must help them by showing the way out and making sure everyone is safe.*

The forbidden acts are, by definition, *unfavorable*, and this is why there is a common interest in curbing or *preventing* them. The permissible acts, for their part, can be categorized depending on how *favorable* they are in terms of their effects on others. Between the classes of forbidden and permissible acts, there lies a class of unfavorable acts. Moral acts contribute to the good of others, often enhancing their well-being.

**Example 9.12**

*A security guard  $S$  assists a customer  $C$  who is looking for a place or an object  $P$  (e.g., a toilet, a garbage bin, the cinema).*

The different modes of social intentionality we have discussed imply certain moral features, as  $A$  is not able to engage itself in any positive cooperation with other agents without refraining from harming them and committing itself to doing its own part in the joint venture. To model the cognitive states and reasoning behind such an action would require a much more detailed BDI architecture than is possible to present within the scope of the present chapter. This might include implementing case-based reasoning in terms of the favorability of the probable outcomes of  $A$ 's actions and a structure of deontic logic covering the concepts of obligation, permission, and forbidden (see Honarvar & Ghasem-Aghaei, 2009).

## Modeling and Implementation

Several agent languages and related tool sets are available for modeling and simulating MASes based on BDI agents (Adam & Gaudou, 2015). One of these toolsets is GAMA (Taillandier et al., 2019), a modeling and simulation environment that focuses on spatial modeling where specifications are written using the GAML language. The GAMA platform provides resources for building simulations within the framework of the classic BDI paradigm that is based on the philosophy of action (Bratman, 1987). Due to the provided support for spatial modeling and graphical visualization, we decided to implement our BDI models utilizing GAMA as our execution platform. These features are highly useful when it comes to modeling the shopping mall domain (see Fig. 9.1). Figure 9.1 illustrates the floor plan of a conventional shopping mall containing walkable areas: a lobby, a movie theater, a restaurant, and a toilet.

A model's entities, processes, and activities were formalized in GAML in terms of *agents*, which, in turn, were specified by their *species*, each with their own attributes, actions, and behaviors. An instance of a species

can perform *actions*. The *action* is a function if it can return a value and a procedure if it cannot. A simple example of a procedure is the action of movement to some point:

```
do goto target: {348,391} ;
```

With a function, we assign the returned value to a variable (here referred to as the *point* type):

```
point target <- find_target(arguments) ;
```

The most critical feature of the BDI architecture is a *plan*, which defines an order of statements that are performed to fulfill some intention. Partial plans created at the time of designing can greatly reduce computational complexity (Bordini et al., 2007); hence, we used plans as offered by GAMA, although we do not touch plans in this chapter. The simplest plan in our simulation was wandering within the space limits:

```
species cleaning_person skills: [moving] control:
  simple_bdi {
    plan lets_wander intention: find_litter {
      do wander bounds: free_space ;
    }...}
```

An agent can *perceive* the environment and change its behavior, mental state, social links, and the like on the basis of the knowledge it acquires. Agents can also interact with each other and change each other's attributes and behavior by means of the *ask* statement (see Example 9.13).

To manage time, GAMA operates using three global variables: cycle (an integer incremented by 1 at each step of the simulation), step (the modifiable duration of a simulation step; 1 second by default), and time (the actual time since the beginning).

To specify the examples described in the chapter, we first described our actions. We first identified the participants, preconditions, and possible additions and deletions. The GAMA implementation was required to follow the rules we had defined and act as an executable specification. Example 9.13 continues from the case of the security guard on a lunch break presented in Example 9.8.

### Example 9.13

*Customer  $C_1$  is next to customer  $C_2$  and notices that  $C_2$  is smoking, that is,  $\text{smoking}(C_2)$  is true, in a location  $L$ . Customer  $C_1$  stores this information as a belief in addition to the location of  $C_2$ . Customer  $C_1$  also develops an intention of sharing information with a guard in we-mode.*

**Action**  $\text{observe\_smoking}(C_1, C_2, L)$ :

- *Preconditions:*  $\text{next}(C_1, C_2)$ ,  $\text{smoking}(C_2)$ .
- *Addition(s):*  $\mathcal{B}_{C_1}(\text{smoking}(C_2))$ ,  $\mathcal{B}_{C_1}(\text{in}(C_2, L))$ ,  
 $\mathcal{I}_{C_1}(\text{give\_information}(S))$ ,  $\text{mode}(C_1, \text{give\_information}(S), \text{wm})$ .

*Security guard  $S$  needs to be in progroup I-mode or we-mode to help a customer  $C$ . Thus, it may happen that  $S$  enters we-mode (and commits to the intention to guard the mall, depicted by the predicate  $\text{patrol}(S)$ ) when  $S$  and  $C$  come next to one another if  $S$  is not in that mode at the time. In some cases,  $S$  may not be able to enter progroup I-mode or we-mode, and, thus, will not be able to help  $C$ . This alternate case is omitted here.*

**Action**  $\text{commit\_to\_help\_after\_approach}(C, S)$ :

- *Preconditions:*  $\text{next}(C, S)$ ,  $\mathcal{D}_S(\text{patrol}(S))$ ,  $\neg\mathcal{I}_S(\text{patrol}(S))$ .
- *Addition(s):*  $\mathcal{I}_S(\text{patrol}(S))$ ,  $\text{mode}(S, \text{patrol}(S), \text{wm})$ .

Customer  $C_1$  informs security guard  $S$  about customer  $C_2$  smoking in location  $L$ .

**Action**  $\text{inform\_of\_smoking}(C_1, C_2, S, L)$ :

- **Preconditions:**  $\text{mode}(S, \text{patrol}(S), \text{wm}), \mathcal{B}_{C_1}(\text{smoking}(C_2)), \mathcal{B}_{C_1}(\text{in}(C_2, L)), \text{next}(C_1, S), \mathcal{I}_{C_1}(\text{give\_information}(S))$ .
- **Additions(s):**  $\mathcal{B}_S(\text{smoking}(C_2)), \mathcal{B}_S(\text{in}(C_2, L))$ .
- **Deletion(s):**  $\mathcal{I}_{C_1}(\text{give\_information}(S))$ .

*In the GAMA implementation, each customer is observing the area within its viewing distance. If customer  $C_1$  notices another customer  $C_2$  smoking (checking the Boolean `smoking` feature),  $C_1$  obtains a new belief containing  $C_2$ 's location and develops a desire to approach the security  $S$  (if such a desire is not already present in  $C_1$ ). Customer  $C_1$  approaches  $S$ , and if the latter is in progroup *I-mode* or *we-mode*,  $C_1$  shares its belief about the smoker's location within the `inform_security` plan. If  $S$  is in pure *I-mode*,  $C_1$  first attracts  $S$ 's attention and asks to receive the `react_to_customer` intention.  $S$  then decides whether he wants to abandon the pure *I-mode* and listen to the customer or not. With the probability of 50%,  $S$  shows that he is ready to be informed and asks the customer to proceed with the `inform_security` plan. Otherwise,  $S$  stops reacting to the customer, and the latter abandons the `approach_security` plan but does not receive an intention to share his knowledge. For the sake of brevity, the code below has been simplified.*

```

species customer ... {
...
  perceive target: customer in: view_dist {
    if self.smoking = true and myself.get_desire_with_name
      (name: "approach security") = nil {
      add_belief(new_predicate(smoking,
        ["location":self.location]));
      do add_desire(approach_security);}}

  plan approach_security_plan intention: approach_security {
    if (location distance_to security < size) {
      // if the security is close enough
      if security.mode = "pure I-mode" {
        ask security {
          do add_intention(react_to_customer);
          focus_customer <- myself.name ;
        }
        do remove_intention(approach_security);
      } else {
        remove_intention(approach_security);
        do add_intention(inform_security);
      }
    } else {
      do goto: security}}

  plan inform_security_plan intention: inform_security {
    smoking_location <- ... // extraction of the
      coordinates from the focus_customer's belief base
    ask security {
      do add_belief(new_predicate(smoking,["location":
        smoking_location]));
    }
    do remove_belief(new_predicate(smoking,["location":
      smoking_location]));
    remove_intention(inform_security);
  }}
}}

species security ... {
  plan react_to_customer_plan intention: react_to_customer {
    if flip(0.5) { // the probability of reaction is 50%
      mode <- "we-mode";
      do add_intention(patrol);
      ask current_customer {
        do add_intention(inform_security) ;
      }
      do remove_intention(react_to_customer, false);
    } else {
      do remove_intention(react_to_customer, false);
    }
  }}
}}

```

## Related Research

Bosse et al. (2011) created a model for describing the reasoning process of other agents utilizing the BDI concepts, namely, beliefs, desires, and intentions, and the theory of mind. Norling (2004) utilized BDI features that resemble folk psychology to incorporate psychological abilities such as knowledge acquisition and decision-making into agent modeling. Adam et al. (2009) proposed a logical formalization to embed emotions into agent models. Cranefield and Dignum (2019) suggested a way to integrate social aspects into BDI agent systems by modeling social practices.

To inhibit unwanted outcomes of actions, there must be constraints in place that rule out as many of such consequences as possible. Norms are deontic statements that are employed to define which (types of) desires and intentions  $A$  must not try to realize through actions. Traditional approaches to reasoning pertaining to norms are based on modal logic (Garson, 2021) and, in particular, *deontic logic*, which can be utilized to formalize obligations and permissions concerning conditions, in analogy to using modal operators in the description of BDI systems.

Criado et al. (2010) extended BDI concepts to model agents that can make pragmatic, autonomous decisions by considering which norms to follow and how to apply them. Such extensions are possible in our approach, enriching the selection of conditions available for modeling. The same can be stated about aspects of time (see, e.g., Urlings et al., 2006) and temporal operators, since obligations and their fulfillment have implications for the past and future.

When considering agent functionality in general, the ability to construct *plans* for the realization of goals and intentions is central, and the same holds true in the context of BDI systems (see, e.g., de Silva et al., 2009; Sardiña et al., 2006) for the *hierarchical* case. Since our approach is compatible with the traditional STRIPS-style planning (Fikes & Nilsson, 1971), we may cover scenarios involving concrete planning or related *verification* tasks. However, for the time being, we have concentrated more on reflexive agents and their use in simulations. A related concept is *crowd simulation* (Cho et al., 2008) that is also relevant to the shopping mall domain but beyond the purview of our focus for now.

## Discussion and Conclusions

This chapter approaches the social dimension of actions performed by agents in terms of modes of social intentionality. The three modes, namely, *pure I-mode*, *progroup I-mode*, and *we-mode*, characterize the interacting agent's intention toward engaging in social relationships with other agents that are relevant to the intended goal and the action being performed. The modes can be applied in various ways in the analysis, definition, and implementation of MASes. First of all, they can be used *implicitly* when modeling actions to understand their true nature and to ease their formalization in general.

The models produced provide possibilities for analyzing, verifying, and simulating agents' behavior. If modes are *explicitly* introduced as variables or conditions in modeling, then a more refined control over execution is enabled via the preconditions of actions. In addition, actions may also manipulate modes as needed if the agents' social intentions change over time, for example, as reactions to other agent's actions or events occurring in the environment. The three modes allow the analysis of positive instances of social action but do not lend themselves to model actions that are disruptive in terms of cooperation as such. In this respect, new conditions of intentionality could be taken into consideration as potential extensions of Tuomela's research (Tuomela, 2007).

Our chapter has, to some extent, been constrained by the limitations of the BDI model itself, which focuses on the three modalities involved, and there is no straightforward way to express the three modes of social intentionality with them. Rather, it was deemed necessary to incorporate modes as factual truths in terms of fluents (cf. the *mode/3* predicate) as part of the agents' states. In reality, agents have much more complex desires and social intentions that can be realized in a number of different ways, each of which could be modeled as a separate plan that further comprises steps involving intentions. Such a recursive structure seems extensive, but without it, a large amount of the specification moves to program code. A major step in our future work will be to tackle these limitations. There are also notable aspects in modeling that have been left unaddressed and will be considered in future work. Most importantly,

the progroup-I-mode and the we-mode presume a group of peer agents. The group dynamics (forming and maintaining groups) and premises for trust are complicated issues in themselves that warrant further attention in the future.

**Acknowledgments** All co-authors of this chapter have been partially supported by the Academy of Finland's Strategic Research Council funded project *Ethical AI for the Governance of Society* (ETAİROS, grant #352441).

## References

- Adam, C., & Gaudou, B. (2015). *BDI agents in social simulations: A survey* (No. RR-LIG-050, LIG). Les rapports de recherche du Laboratoire d'Informatique de Grenoble.
- Adam, C., Herzig, A., & Longin, D. (2009). A logical formalization of the OCC theory of emotions. *Synthese*, 168(2), 201–248. <https://doi.org/10.1007/s11229-009-9460-9>
- Balaji, P. G., & Srinivasan, D. (2010). An introduction to multi-agent systems. In D. Srinivasan & L. C. Jain (Eds.), *Innovations in multi-agent systems and applications—1* (pp. 1–27). Springer. [https://doi.org/10.1007/978-3-642-14435-6\\_1](https://doi.org/10.1007/978-3-642-14435-6_1)
- Bordini, R. H., Hübner, J. F., & Wooldridge, M. (2007). *Programming multi-agent systems in AgentSpeak using Jason*. John Wiley & Sons, Inc.
- Bosse, T., Memon, Z. A., & Treur, J. (2011). A recursive BDI agent model for theory of mind and its applications. *Applied Artificial Intelligence*, 25(1), 1–44.
- Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.
- Broersen, J. M., Dastani, M., Hulstijn, J., Huang, Z., & van der Torre, L. W. N. (2001). The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In E. Andre, S. Sen, C. Frasson, & J. P. Mueller (Eds.), *Proceedings of AGENTS 2001* (pp. 9–16). ACM. <https://doi.org/10.1145/375735.375766>
- Cho, K., Ikerani, N., Kikuchi, M., Nishimura, K., Hayashi, H., & Hattori, M. (2008). BDI model-based crowd simulation. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *Proceedings of IVA 2008* (pp. 364–371). Springer. [https://doi.org/10.1007/978-3-540-85483-8\\_37](https://doi.org/10.1007/978-3-540-85483-8_37)

- Cranefield, S., & Dignum, F. (2019). Incorporating social practices in BDI agent systems. In E. Elkind, M. Veloso, N. Agmon, & M. E. Taylor (Eds.), *Proceedings of AAMAS 2019* (pp. 1901–1903). IFAAMAS.
- Criado, N., Argente, E., & Bortti, V. J. (2010). A BDI architecture for normative decision making. In W. van der Hoek, G. A. Kaminka, Y. Lesperance, M. Luck, & S. Sen (Eds.), *Proceedings of AAMAS 2010* (pp. 1383–1384). IFAAMAS.
- de Silva, L., Sardiña, S., & Padgham, L. (2009). First principles planning in BDI systems. In C. Sierra, C. Castelfranchi, K. S. Decker, & J. S. Sichman (Eds.), *Proceedings of AAMAS* (Vol. 2, pp. 1105–1112). IFAAMAS.
- Fikes, R. E., & Nilsson, N. J. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. In D. C. Cooper (Ed.), *Proceedings of IJCAI'71* (pp. 608–620). Morgan Kaufmann.
- Garson, J. (2021). Modal logic. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021 ed.). Stanford University.
- Gelfond, M., & Lifschitz, V. (1998). Action languages. *Electronic Transactions on Artificial Intelligence*, 2, 193–210. <http://www.ep.liu.se/ej/etai/1998/007/>
- Hallamaa, J., & Kalliokoski, T. (2020). How AI systems challenge the conditions of moral agency? In M. Rauterberg (Ed.), *Proceedings of culture and computing, C&C 2020* (pp. 54–64). Springer. [https://doi.org/10.1007/978-3-030-50267-6\\_5](https://doi.org/10.1007/978-3-030-50267-6_5)
- Honarvar, A. R., & Ghasem-Aghaee, N. (2009). Casuist BDI-agent: A new extended BDI architecture with the capability of ethical reasoning. In H. Deng, L. Wang, F. L. Wang, & J. Lei (Eds.), *Artificial intelligence and computational intelligence* (pp. 86–95). Springer. [https://doi.org/10.1007/978-3-642-05253-8\\_10](https://doi.org/10.1007/978-3-642-05253-8_10)
- Koster, A., Schorlemmer, W. M., & Sabater-Mir, J. (2013). Opening the black box of trust: Reasoning about trust models in a BDI agent. *Journal of Logic and Computation*, 23(1), 25–58. <https://doi.org/10.1093/logcom/exs003>
- Labrou, Y., & Finin, T. (1994). A semantics approach for KQML—A general purpose communication language for software agents. In *Proceedings CIKM'94* (pp. 447–455). ACM.
- Neumann, M. (2010). Norm internalisation in human and artificial intelligence. *Journal of Artificial Societies and Social Simulation*, 13(1). <https://doi.org/10.18564/jasss.1582>
- Norling, E. (2004). Folk psychology for human modelling: Extending the BDI paradigm. In *Proceedings of AAMAS 2004* (pp. 202–209). IEEE Computer Society.

- Rao, A., & Georgeff, M. (1991). Modeling rational agents within a BDI architecture. In J. F. Allen, R. Fikes, & E. Sandewall (Eds.), *Proceedings of KR 1991* (pp. 473–484). Morgan Kaufmann.
- Rocha, J., Boavida-Portugal, I., & Gomes, E. (2017). Introductory chapter: Multi-agent systems. In J. Rocha (Ed.), *Multi-agent systems* (pp. 3–13). IntechOpen.
- Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Sardiña, S., de Silva, L., & Padgham, L. (2006). Hierarchical planning in BDI agent programming languages: A formal approach. In H. Nakashima, M. P. Wellman, G. Weiss, & P. Stone (Eds.), *Proceedings of AAMAS 2006* (pp. 1001–1008). ACM.
- Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence*, 60(1), 51–92. [https://doi.org/10.1016/0004-3702\(93\)90034-9](https://doi.org/10.1016/0004-3702(93)90034-9)
- Taillandier, P., Gaudou, B., Grignard, A., Huynh, Q.-N., Marilleau, N., Caillou, P., Philippon, D., & Drogoul, A. (2019). Building, composing and experimenting complex spatial models with the GAMA platform. *GeoInformatica*, 23(2), 299–322. <https://doi.org/10.1007/s10707-018-00339-6>
- Tuomela, R. (2007). *The philosophy of sociality. The shared point of view*. Oxford University Press.
- Urlings, P., Sioutis, C., Tweedale, J., Ichalkaranje, N., & Jain, L. C. (2006). A future framework for interfacing BDI agents in a real-time teaming environment. *Journal of Network and Computer Applications*, 29(2–3), 105–123. <https://doi.org/10.1016/j.jnca.2004.10.005>
- von Wright, G. H. (1968). *An essay in deontic logic and the general theory of action*. North-Holland Pub. Co.
- Woolridge, M. (2009). *Introduction to multi-agent systems* (2nd ed.). Wiley.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

