

# A Formal Framework for Designing Boundedly Rational Agents

Andreas Brännström<sup>1</sup><sup>a</sup>, Timotheus Kampik<sup>1</sup><sup>b</sup>, Ramon Ruiz-Dolz<sup>2</sup><sup>c</sup> and Joaquin Taverne<sup>2</sup><sup>d</sup>

<sup>1</sup>Umeå University, Umeå, Sweden

<sup>2</sup>Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Valencia, Spain

**Keywords:** Rational Agents, Socially Intelligent Systems, Engineering Multi-agent Systems.

**Abstract:** Notions of *rationality* and *bounded rationality* play important roles in research on the design and implementation of autonomous agents and multi-agent systems, for example in the context of instilling socially intelligent behavior into computing systems. However, the (formal) connection between artificial intelligence research on the design and implementation of boundedly rational and socially intelligent agents on the one hand and formal economic rationality – *i.e.*, choice with clear and consistent preferences – or instrumental rationality – *i.e.*, the maximization of a performance measure given an agent’s knowledge – on the other hand is weak. In this paper we address this shortcoming by introducing a formal framework for designing boundedly rational agents that systematically relax instrumental rationality, and we propose a system architecture for implementing such agents.


## 1 INTRODUCTION


In the Artificial Intelligence (AI) research community, a key line of research is concerned with the design and implementation of *rational agents*, *i.e.*, agents that work towards *goals* they *intend* to achieve under consideration of the *beliefs* they hold about their environment (Wooldridge, 1997). To advance engineering perspectives on this line of research (and to facilitate application potential), a broad variety of tools and frameworks have been introduced over the past two decades (Kravari and Bassiliades, 2015); a notable example that enjoys popularity is Jason (Bordini et al., 2007) (as well as the JaCaMo framework (Boissier et al., 2013) that make use of Jason), an interpreter of a dialect of the *AgentSpeak* programming language (Rao, 1996), which combines logic and agent-oriented programming paradigms. Generally, the state of the art of research and development of tools and frameworks for implementing *rational agents* is well-surveyed (Cardoso and Ferrando, 2021; Kravari and Bassiliades, 2015), and the community frequently reflects on shortcomings and ways to address them (Mascardi et al., 2019; Logan, 2018).


While the notion of *rationality* that the AI research community has is notably broader than the precise formal properties of rationality that are a cornerstone


of economic theory (Osborne and Rubinstein, 2020), a central element is *goal-orientation*, *i.e.*, the ability of an agent to maximize goal attainment, while potentially compromising between several conflicting goals. Analogously to how formal models of bounded rationality like Tversky’s and Kahneman’s *prospect theory* (Kahneman and Tversky, 1979) systematically relax the properties (in particular the maximization of expected utility), some models, tools, and frameworks for designing and implementing intelligent agents try to relax the rationality constraints of these agents. However, these approaches are typically not grounded in a formal model of rationality or in a systematic relaxation thereof: a generic, abstract formal framework for boundedly rational agents that can serve as a point of departure does not exist. This paper works towards addressing this issue by answering the following questions:

1. What notions of rationality and bounded rationality exist in the literature, and how are these notions reflected by practical approaches to engineering boundedly rational agents and multi-agent systems?
2. How can we create a more precise formal framework of a boundedly rational agent as a point of departure for engineering efforts?
3. Based on this formal framework, how can we devise a holistic, engineering-centered, yet technology-agnostic meta-model for the design and implementation of boundedly rational agents?

<sup>a</sup> <https://orcid.org/0000-0001-9379-4281>

<sup>b</sup> <https://orcid.org/0000-0002-6458-2252>

<sup>c</sup> <https://orcid.org/0000-0002-3059-8520>

<sup>d</sup> <https://orcid.org/0000-0002-5163-5335>

The rest of this paper is organized as follows. Section 2 provides an overview of different notions of rationality and bounded rationality to then proceed with a survey of approaches and frameworks for designing and implementing boundedly rational agents. Section 3 addresses some shortcomings of existing research by introducing a formal framework of boundedly rational agents and by proposing a meta-model (technology-agnostic architecture) therefor. Section 4 then discusses how the formal framework and architectural meta-model can be integrated with existing approaches to designing and implementing somewhat boundedly rational agents, before Section 5 concludes the paper.

## 2 DESIGNING AND IMPLEMENTING BOUNDEDLY RATIONAL AGENTS

In this section, we first provide a formal definition of the notions of rationality and bounded rationality to then give an overview of existing approaches to designing and implementing boundedly rational agents.

### 2.1 Rational Agents

The notion of a rational actor or *agent* comes from economic rationality and has a well-defined formal meaning (see, e.g. (Osborne and Rubinstein, 2020)). Given a set of choice options  $A$ , a rational agent's choice selects an element  $a^* \in A$ , which then establishes a partial order<sup>1</sup>  $\succeq$  on  $A$ , such that  $\forall a \in A$ ,  $a^* \succeq a$ , i.e.,  $a^*$  is preferred over all (other) possible choices from  $A$ . When the set of choices  $A$  is expanded to  $A'$  (i.e., new elements are added,  $A \subset A'$  or technically  $A \subseteq A'$ ), the new choice of  $a'^* \in A'$  must establish preferences that are consistent with respect to the preferences established by the previous choice, which can be concisely summarized as if  $a'^* \in A$  then  $a'^* = a^*$ , *ceteris paribus* (assuming all else remains the same, i.e., the agent's knowledge of the world did not evolve).

However, in the AI community, economic rationality is less prominent than the notion of rationality as agent behavior that strives to maximize a performance measure, which is often not formally defined, but which boils down to the maximization of expected utility, given the knowledge at hand (and

<sup>1</sup>Recall that a partial order  $\succeq$  on a set  $S$  is reflexive, anti-symmetric, and transitive, i.e.,  $\forall a, b, c \in S$ , the following statements hold true: i)  $a \succeq a$ ; ii) if  $a \succeq b$  and  $b \succeq a$  then  $a = b$ ; iii) if  $a \succeq b$  and  $b \succeq c$  then  $a \succeq c$ .

hence can be traced back to Von Neumann's and Morgenstern's utility theory). To separate these two notions of rationality, Gintis distinguishes between *formal (economic) rationality* and *instrumental rationality* (Gintis, 2018). Beyond these two notions, some symbolic AI researchers formally define rationality in the form of *rationality postulates*, e.g. for belief revision (Alchourrón et al., 1985) and formal argumentation (Caminada, 2017). In these cases, entirely new notions of rationality have been introduced that rely on the authors' intuitions and are unrelated to economic and instrumental rationality. To conclude, the notion of *rationality* is ambiguous in symbolic artificial intelligence research, which may render the systematic treatment of *bounded rationality* a challenge as well.

### 2.2 Boundedly Rational Agents

From an engineering perspective, a boundedly rational agent is an agent whose "rationality degree" (colloquially speaking) has been deliberately and systematically relaxed. Depending on the requirements of the application domain, an agent implementation can benefit from the relaxation of pure economic or instrumental rationality. For example, when interacting with humans, the integration of the emotional aspect of human nature into an agent's behavior can improve human perception of AI. This "*humanization*" of intelligent agents' reasoning is reflected in the three major components of a boundedly rational agent: (i) the affective/emotional modeling (e.g., social cognition (Seyfarth and Cheney, 2015)), (ii) the explainability of an agent's decisions, and (iii) the Human-Aware Planning (Chakraborti, 2018). All these three components are covered under the umbrella of Computational Argumentation, the research area that studies the integration of human (argumentative) reasoning into an intelligent system (e.g., agent), making it an interesting and powerful approach to undertake the implementation of a complete boundedly rational agent. Thus, prior to creating our formal framework and architectural meta-model for implementing boundedly rational agents, we review the most prominent existing works on affective/emotional agents, explainable agents, human-aware planning techniques, and computational argumentation approaches.

As discussed above, in the agent-oriented computing paradigm, traditional proposals have generally focused on economic theory or Aristotelian practical reasoning to define the behavior of agents (Wooldridge, 1997). However, these approaches have certain disadvantages when it comes to modeling human social or affective behavior as

well as when modeling human social organizations. This is because human beings are not always following a logical reasoning process, aimed at maximizing or minimizing their benefits, but rather follow a reasoning process that is biased in part by emotions, social context, or other individual characteristics such as personality, gender, or age (Segerstrom and Smith, 2019). In humans, these factors are closely related to other cognitive processes commonly used in the field of artificial intelligence such as reasoning, planning or decision making processes (Davis et al., 2007; Grecucci et al., 2020). Considering this, a natural approach to the concept of bounded rational agents can be found in the field of affective computing (Picard, 1997). Affective computing studies the modeling of affective and social factors to improve models for representing, processing, understanding, and simulating human affective behavior. Most of the models proposed in this field try to simulate the influence of cognitive processes related to the agent's affective abilities on deliberation processes (Gebhard, 2005; Taverner et al., 2021; Paiva et al., 2017). Thus, emotions and affective characteristics are used to relax the level of rationality of the agents to reach a more human-like level of affective behavior simulation. For example, in (Dias et al., 2014), the *FAtiMA* architecture is presented. This architecture is based on the Belief-Desire-Intention (BDI) model and uses different affective factors, such as emotions or personality, to influence the cognitive reasoning process of the agent. To select the emotion according to the agent's situational context, a model based on appraisal theories is used (Ojha and Williams, 2017). Once the emotion is determined, the agent's decision making processes is altered, influencing agent's rational processes. Similarly, *ALMA*, A Layered Model of Affect, is described in (Gebhard, 2005). The authors estimate the agent's mood when an event occurs and then use it to determine the action the agent should execute. Another interesting approach is the one described in (Alfonso et al., 2017), in which the *GenIA*<sup>3</sup> architecture is presented. *GenIA*<sup>3</sup> is a general-purpose architecture that extends the AgentSpeak language for the development of affective agents. That architecture combines both affective (based on different appraisal, cognitive, and affective theories) and rational deliberation processes (based on the BDI architecture). The affective responses elicited when appraising events influence the agent's inference and decision making processes (considering factors such as mood, emotions, expectations, or personality) to a greater or lesser extent depending on a parameter that determines the level of rationality of the agent. *GenIA*<sup>3</sup> is currently being implemented as an extension of Jason and is developed

using a modular design that allows for the adaptation to other emotion theories.

Recently, research on *explainable* agents and multi-agent systems has emerged as a high-profile topic within the community during the last years (Anjomshoae et al., 2019)<sup>2</sup>. To facilitate human-agent explainability, one key idea is to design artificial agents whose deliberation processes resemble – to some limited extent – human reasoning and decision-making (Broekens et al., 2010). More broadly speaking, explainable artificial intelligence is expected to rely on social science insights as a necessary requirement for long-term break-throughs (Miller, 2019). This assumption is reflected in a range of works that *i)* study agent explainability from a behavioral psychology/behavioral economics perspective (Kampik et al., 2019a; Tulli et al., 2019) and *ii)* apply cognitive architectures, such as architectures based on the Belief-Desire-Intention deliberation approach, that have their roots in cognitive science, psychology, or philosophy to empirically study explainability in human-agent/human-MAS interaction (Mualla et al., 2021; Broekens et al., 2010).

On the formal side, *machine reasoning explainability* is commonly considered a formal property (or set thereof) that allows to provide a set of explanations (e.g., beliefs in an agent's belief base) of why a certain decision has been made or a certain inference has been drawn (Čyras et al., 2020). However, symbolic reasoners and planners are typically considered *rational*. For instance, in the context of formal argumentation (see the previous subsection, as well as (Čyras et al., 2021)), the argumentation approaches that are to be explained are typically considered *rational*; no explanations as to why an agent is merely *boundedly rational* – i.e., why rationality is violated – need to be provided. Still, to facilitate explainability, or more precisely: human interpretability, a symbolic reasoner or planner will adjust its behavior so that it is easier to explain to a human user, or better aligns with human intuition without an explanation being necessary, an empirical characteristic that is often referred to as *explicability* (Zhang et al., 2017). Therefore, the overlap of explainable and boundedly rational agents is the need for human-like decision-making to facilitate explainability and explicability in the context of *human-aware planning*.

Human-aware planning is an algorithmic approach for an autonomous system to plan its actions to cohabit in an environment that is populated and/or affected by humans (Chakraborti, 2018). This requires a system to estimate what future actions hu-

<sup>2</sup>Note that we consider the domain of explainable machine learning as out of the scope of this work.

mans might take in that environment (Cirillo et al., 2010). In order to achieve shared and personal goals, a system needs ways to recognize the human's plan (sequence of actions) and goal, and align its own plan (another sequence of actions) with the human's plan. This can mean to not go for the optimal plan in terms of traditional efficiency measures (e.g., time, shortest route, etc.), but a sub-optimal plan that aligns with human behavior and reasoning.

The human-aware planning problem has, in general, been explored in scenarios where a robot is situated in an environment involving humans, where the robot perceives the human through sensors and, through a model of human behavior, it adjusts its deliberative (planning) process. This produces a merged plan where the robot adapts its plan to comply with the constraints of the human plan (Chakraborti et al., 2018; Cirillo et al., 2010). For example, the work proposed in (Köckemann et al., 2014), addresses the challenge of automatically generating plans that have to accommodate scheduled activities, features and preferences of humans. The planning algorithm uses causal reasoning to create a plan using heuristic forward planning together with a causal graph (Helmert and Geffner, 2008). Another work (Floyd et al., 2018) explores goal reasoning agents that are able to dynamically reason about their goals, and modify them in response to unexpected events or opportunities. The approach allows for agents that are members of human-agent teams to use the partially specified preferences of the human to estimate the utility of goals and guide goal selection. The work utilizes the SapaReplan framework (Talamadupula et al., 2010).

In contrast to typical applications of human-aware planning where a robot is situated in an environment populated by humans, the research conducted in (Brännström et al., 2020), explores how a software agent can influence a human's behavior by adapting aspects of the human's environment, introducing HA-TPB, a human-aware planning architecture based on the theory of planned behavior (TPB) (Ajzen et al., 1991). According to TPB, a mental model can be derived from three sources of human beliefs (attitude, subjective norm, and perceived behavior control) which are linked to motivation, intention and goals. The HA-TPB architecture captures the casual relation between a human's beliefs and behavior by a transition system modeled in action reasoning (Gelfond and Lifschitz, 1998) to deliberate about the human's behavior. An example use-case of the HA-TPB architecture is a Virtual Reality (VR) game, in which an agent is used for providing assistance in a social scenario to children with autism by adapting the virtual environment. The software agent evaluates the

human's plan and adapts its actions by considering human reasoning. In the case of autism, this requires the system to understand the child's limitations and what assistance the child may need, in order to perform a wanted behavior successfully. This can be seen as a type of relaxation of the agent's rationality, since the agent does not follow a classical reasoning pattern aimed at achieving a goal, but rather the agent modifies its behavior to adapt it to the behavior of its human interlocutor.

Finally, the idea of bounded rationality has also been studied and analyzed from different perspectives within computational argumentation theory. Computational argumentation is one of the main branches of AI that explores the integration of rationality into computer systems through the use of arguments and argumentation semantics (Dung, 1995; Atkinson et al., 2017; Ruiz-Dolz, 2020). However, as we have discussed above, human behavior does not always follow a rational argumentative pattern, but in most cases is guided by other affective or social factors.

Still, from a technical perspective, most (abstract) argumentation semantics (Baroni et al., 2011) do not satisfy the *consistent preferences* principle of economic rationality, and the systematic treatment of economic rationality and bounded rationality in the context of computational argumentation is an emerging research frontier (Kampik and Gabbay, 2021). In this context, the relation between systematically relaxing (economic) rationality and the systematic relaxation of monotony of entailment can be considered of particular interest.

Several approaches explore how bounded rationality can be integrated within the frame of formal computational argumentation research. The main objective of such approaches is the relaxation of the notion of acceptance for an argument (i.e., when an argument can be considered valid or not from an argumentative viewpoint). The main differences between the observed approaches can be found in the reasoning aspect that it is relaxed in the definition of acceptance. For example, in Defeasible Logic Programming (DeLP) argumentation (García and Simari, 2004; Pollock, 1987), the authors propose an alternative paradigm for computational argumentation where arguments may be brought into consideration even if their deductive validity is not provided (but need to be rationally compelling). Thus, DeLP-based argumentative approaches relax the internal reasoning aspect of argument structures so that they can be used in less informed environments. A different theoretical approach for bounded rationality in computational argumentation was introduced with the definition of (epistemic) probabilistic argumentation frame-

works (Thimm, 2012). Combining argumentation frameworks and probabilistic reasoning, these probabilistic frameworks include uncertain information related to the credibility of arguments in their formalization. In an argumentative dialogue, it is hard to model whether an argument can be believed or not. Hence, these frameworks define the acceptance of an argument based on their expected believability (i.e., a probability distribution) rather than relying uniquely on purely rational aspects. Finally, an important viewpoint on bounded rationality for the theory of computational argumentation is the consideration of human mental properties, such as emotions. Emotions are a characteristic feature of humans, where emotional and rational behaviors coexist. Human rationality is usually influenced by the activation of emotions, so it can be an important dimension to be brought into consideration when defining computational models of bounded rationality. Emotional argumentation frameworks (Dalibón et al., 2012) integrate the emotional aspect into the *abstract argumentation* (Dung, 1995) theory. Furthermore, within this approach, emotions are taken into account during the formal evaluation of argumentation. Thus, an emotional state can influence the acceptability of an argument under this paradigm. In brief, bounded rationality in computational argumentation has been theoretically explored from the point of view of the relaxation of the acceptability notion (e.g., regarding the internal structure of arguments, credibility/believability, or emotional aspects of humans).

Different implementations of argumentation-based systems and agents have been applied to various domains: for assisting with privacy management in online social networks (Kökciyan et al., 2017; Ruiz-Dolz et al., 2019); for automatically generating explanations and recommendations (Cocarascu et al., 2019; Heras et al., 2020); for assisting with negotiation protocols (Amgoud et al., 2007; de Jonge and Sierra, 2017; Bouslama et al., 2020); and for healthy eating assistance (Thomas et al., 2019) among others. Some of these works have been complemented with field studies that analyse the variations on the perceived strength of an argument depending on non-rational human features (e.g., personality or social features) (Thomas et al., 2017; Ciocarlan et al., 2019; Ruiz-Dolz et al., 2021). Despite these efforts, we were not able to identify any work focusing exclusively on the implementation of an argumentation-based boundedly rational agent or system. Furthermore, many tools and libraries for argumentation-based reasoning exist (Alviano, 2017; Cerutti et al., 2016a; Cerutti et al., 2016b; Craandijk and Bex, 2020), but bounded rationality is not thor-

oughly brought into consideration.

In this section, different approaches to bounded rationality in agents and AI have been reviewed, and a classification of the identified approaches into the major components of a boundedly rational agent has been provided. These components define the three pillars that group most of previous research in bounded rationality from the AI viewpoint. However, we observe that no generic frameworks for designing and implementing boundedly rational agents exist, neither formal frameworks, nor architectural meta-models. Most of the reviewed research focuses on a very specific aspect or domain (e.g., affective computing or human assistance), ignoring the formal definitions of economic and instrumental rationality, and without considering their general implications for bounded rationality in AI.

### 3 FORMAL FRAMEWORK AND ARCHITECTURAL META-MODEL

To address two of the shortcomings identified in Section 2, i.e., the lack of generic formal and architectural frameworks for bounded *instrumental* rationality, this Section first introduces an abstract formal framework for modeling boundedly rational agents, and then an architectural meta-model for implementing them.

#### 3.1 Formal Framework

Let us introduce an abstract, generic formal framework for a boundedly rational agent. As a prerequisite, we introduce a (rational) *agent function*, which in turn maximizes an expected utility function, given a specific *percept sequence*, in which sets of percepts are typically temporally ordered.

**Definition 1** (Percept Sequence). *A percept sequence  $S$  is a sequence  $\langle P_0, \dots, P_t \rangle$ , where for  $0 \leq i \leq t$ ,  $P_i$  is a set of elements (which we call “percepts”).*

Let us introduce the expected utility function, which given a percept sequence and an action  $a$ , returns an agent’s expected utility of this action. In this context, we consider an action a logical literal.

**Definition 2** (Expected Utility Function). *Let  $S$  be a set of percept sequences (our percept sequence space) and let  $\mathcal{A}$  be a set actions (our action space). The expected utility function  $u : S \times \mathcal{A} \rightarrow \mathbb{R}$  takes a percept sequence  $S \in S$  and an action  $a \in \mathcal{A}$  and returns the action’s expected utility  $u_e \in \mathbb{R}$  given the percept sequence.*

Note that for the sake of simplicity, we assume the expected utility of an action is represented as a real number. However, we concede that other representations, like rankings in a preference order, are possible and indeed preferable in some contexts.

Now, we can define the rational agent function.

**Definition 3** (Rational Agent Function). *Let  $S$  be a set of percept sequences, let  $\mathcal{A}$  be a set of actions, and let  $u$  be an expected utility function. The rational agent function  $f^u : S \rightarrow \mathcal{A}$  takes a percept sequence  $S \in \mathcal{S}$  and returns an action  $a \in \mathcal{A}$ , such that  $a \in \arg \max_{a' \in \mathcal{A}} u(a', S)$ .*

Given a utility function  $u$  and a rational agent function  $f_u : S \rightarrow \mathcal{A}$ , we call  $S$  the *percept sequence space* of  $f_u$  and  $\mathcal{A}$  the *action space* of  $f_u$ .

To allow for the specification of boundedly rational agent, we extend the rational agent function.

**Definition 4** (Boundedly Rational Agent Function). *Let  $S$  be a set of percept sequences, let  $\mathcal{A}$  be a set of actions, let  $u$  be an expected utility function, and let  $f^u$  be a rational agent function. The boundedly rational agent function  $g_{f_u, p, q} : S \rightarrow \mathcal{A}$  takes a percept sequence  $S \in \mathcal{S}$  and returns an action  $a \in \mathcal{A}$ , such that for every  $S \in \mathcal{S}$ , it holds that  $g_{f_u, p, q}(S) = q(f_u(p(S)), p(S))$ , where:*

- $p : S \rightarrow \mathcal{S}$  is the percept sequence pre-processing function;
- $q : \mathcal{A} \times \mathcal{S} \rightarrow \mathcal{A}$  is the action post-processing function.

Let us introduce three simplistic examples that illustrate how the boundedly rational agent function can be applied. For all examples, we assume an action space  $\mathcal{A}$ , a percept sequence space  $\mathcal{S}$ , an expected utility function  $u_e : S \times \mathcal{A} \rightarrow \mathbb{R}$ , and a rational agent function  $f^u : S \rightarrow \mathcal{A}$ .

**Example 1: Forgetful Agent.** A forgetful boundedly rational agent only considers the most recent percept sequences when deciding on an action. We define the forgetful agent's boundedly rational agent function  $g_{f_u, p, q} : S \rightarrow \mathcal{A}$ , such that for every  $S = \langle P_0, \dots, P_t \rangle \in \mathcal{S}$ ,  $a \in \mathcal{A}$ :

$$p(S) = \begin{cases} \langle P_{t-1}, P_t \rangle, & \text{if } |S| \geq 2; \\ \langle P_t \rangle, & \text{otherwise;} \end{cases}$$

$$q(a, S) = a.$$

**Example 2: Imprecisely Utilitarian Agent.** An imprecisely utilitarian agent always take the action with the “second best” utility, considering its utility function. We define the imprecisely agent's boundedly rational function  $g'_{f_u, p', q'} : S \rightarrow \mathcal{A}$ , such

that for every  $S = \langle P_0, \dots, P_t \rangle \in \mathcal{S}$ ,  $a \in \mathcal{A}$ :

$$p'(S) = S;$$

$$q'(a, S) = \arg \max_{a' \in \mathcal{A} \setminus \{a\}} u(a', S).$$

**Example 3: Forgetful, Imprecisely Utilitarian Agent.** We can combine the forgetful agent and the imprecisely utilitarian agent by defining the boundedly rational agent function  $g''_{f_u, p, q'} : S \rightarrow \mathcal{A}$ .

### 3.2 Architectural Meta-model

Considering the variety of tools and frameworks for implementing somewhat *boundedly* rational agents that have emerged over the years in the literature, we argue that presenting a generic, technology-agnostic architecture, as well as a technology-specific instantiation thereof is valuable. The conceptual *boundedly rational agent architecture* can be described as follows (see Figure 1). As usual, the architecture models the interaction of agents with their environment (and with each other through the environment). In contrast to other prominent conceptual architectures like the JaCaMo meta-model (Boissier et al., 2013), the boundedly rational agent architecture does not explicitly model artifacts and organizations, as they are considered out of scope. Instead, the focus lies on *i*) particular agent internals and *ii*) the novel concept of agent-to-agent cognitive theory discovery.

**Agent Internals.** From an *agent internals* perspective, we split the mind of a boundedly rational agent into two modules: a *rational agent* module that may, for example, implement a classical BDI reasoning loop, and a *boundedly rational* module, that constraints the perception, reasoning steps, and actions of the agent. These modules communicate via an abstraction layer that serves as a middleware between the two modules, in particular in case of technological/implementation-specific differences. In this way, the boundedly rational module allows for the systematic relaxation of rational agent behavior, as defined more precisely by our formal framework.

**Cognitive Theory Discovery.** A boundedly rational agent should be able to make its cognitive theory *discoverable* so that other agents can potentially – if equipped accordingly – interact with the agent in a way that considers the relaxation of rationality. To allow for this, the boundedly rational agent architecture features a discovery module for cognitive theory specifications. For example, when dealing with interactions with human agents, a discovery module can be specified in terms of the

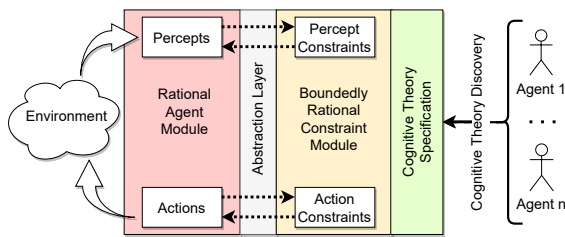


Figure 1: Conceptual architecture for boundedly rational agents.

theory of planned behavior (Ajzen et al., 1991), a cognitive theory that links a human agent’s beliefs to its behavior. The theory manages three sets of beliefs: attitude, subjective norm, and perceived behavioral control, which together can expose a human agent’s behavioral intentions. Still, an agent may choose to not expose all aspects of its cognitive theory to the environment, for example for privacy or cultural reasons; analogously, a human may opt to not disclose their cognitive characteristics, *e.g.*, to avoid an exploitation by malicious parties.

Note that our architecture can be considered as *orthogonal* to rational agents architecture and meta-models, *i.e.*, it abstains from specifying aspects that rational agent architectures already cover (like reasoning-loops) and instead focuses on novel features that are central to boundedly rational agents and to the multi-agent systems in which they (inter)act.

## 4 DISCUSSION

The objective of the formal framework and meta-model that we have introduced in the previous section is to facilitate precise and nuanced perspectives on bounded instrumental rationality when designing and implementing autonomous agents. Research directions that to some extent assume a relaxation of rationality (either in the formal economic or the instrumental sense), *i.e.*, the ones that have been summarized in Section 2, can potentially be integrated with our work.

**Affective Agents.** Human-machine interaction, the simulation of human behavior, or the simulation of human social organizations are some of the areas that can benefit from the advances made in the field of affective computing and affective agents. Models that allow to constrain the level of instrumental rationality of the agents, such as the one proposed in this paper, contribute to improve the simulation of human-like affective behaviors.

**Explainable Agents.** With an increasing demand for transparency/explainability in intelligent systems, a need for techniques to create two-way explanatory interaction systems arises. In this context, the boundedly rational agent framework can provide *i)* an understanding of human behavior to the artificial agent and *ii)* an understanding of the system’s behavior to humans in a human-readable format; in this way, human-AI interaction capabilities can be improved and greater levels of transparency in AI systems can be provided.

**Human-aware Agents.** One goal in the field of human-AI interaction is to create cognitive interactive systems that are human-aware and whose actions and deliberative processes are restricted by interaction constraints that reflect human behavior. In such systems, the boundedly rational agent framework can serve as a foundation for modeling humans. This can provide systems a theory of mind (ToM) (Frith and Frith, 2005) of the human, through which systems can understand and predict human reasoning and behavior, and plan their actions in a human-aware manner (*e.g.*, by considering emotional, motivational and behavioral constraints in their interactions).

**Argumentation-enabled Agents.** Argumentation can be applied to boundedly rational agent internals for belief revision purposes. Also, argumentation can be used to facilitate cooperation between agents whose cognitive theories or instantiations thereof are not fully aligned, for example as outlined in (Kampik et al., 2019b).

Our research can be extended into different directions. *i)* Formally, the framework can be integrated with particular cognitive theories, such as prospect theory (Kahneman and Tversky, 1979) and theory of planned behavior (Ajzen et al., 1991). Also, the abstract framework can be filled with structure, for example using logic-based approaches that treat belief revision as a first-class abstraction. *ii)* From an engineering perspective, reference implementations of the architectural meta-model in different programming languages (agent-oriented or agent-agnostic) can be provided. *iii)* Empirically, the value the framework may provide for particular use cases, for example to facilitate the explainability of boundedly rational agents, can be analyzed.

## 5 CONCLUSION

In this paper, we have introduced a generic, abstract formal framework for designing boundedly rational

agents, as well as an architectural meta-model for implementing such agents in practice, or as engineering research artifacts. These models can potentially facilitate systematic approaches to engineering boundedly rational agents, from both formal foundations-oriented and applied perspectives. Particularly relevant future works are the integration of the abstract formal framework with cognitive theories, for example prospect theory (Kahneman and Tversky, 1979) and the theory of planned behavior (Ajzen et al., 1991), as well as the implementation of the architectural meta-model.

## ACKNOWLEDGEMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the Spanish Government project PID2020-113416RB-I00, and GVA-CEICE project PROMETEO/2018/002.

## REFERENCES

- Ajzen, I. et al. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.
- Alchourrón, C. E., Gärdenfors, P., and Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530.
- Alfonso, B., Vivancos, E., and Botti, V. (2017). Toward formal modeling of affective agents in a BDI architecture. *ACM Transactions on Internet Technology (TOIT)*, 17(1):5.
- Alviano, M. (2017). The pyglaf argumentation reasoner. In *Technical Communications of the 33rd International Conference on Logic Programming, ICLP 2017*, volume 58, pages 2:1–2:3. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.
- Amgoud, L., Dimopoulos, Y., and Moraitis, P. (2007). A general framework for argumentation-based negotiation. In *Argumentation in Multi-Agent Systems, 4th International Workshop, ArgMAS*, volume 4946, pages 1–17. Springer.
- Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19*, page 1078–1088, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.
- Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G. R., Thimm, M., and Villata, S. (2017). Towards artificial argumentation. *AI Mag.*, 38(3):25–36.
- Baroni, P., Caminada, M., and Giacomin, M. (2011). An introduction to argumentation semantics. *Knowl. Eng. Rev.*, 26(4):365–410.
- Boissier, O., Bordini, R. H., Hübner, J. F., Ricci, A., and Santi, A. (2013). Multi-agent oriented programming with jacamo. *Science of Computer Programming*, 78(6):747 – 761.
- Bordini, R. H., Hübner, J. F., and Wooldridge, M. (2007). *Programming Multi-Agent Systems in AgentSpeak Using Jason (Wiley Series in Agent Technology)*. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Bousslama, R., Jordán, J., Heras, S., and Amor, N. B. (2020). Strategies in case-based argumentation-based negotiation: An application for the tourism domain. In *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Trust-worthiness. The PAAMS Collection - International Workshops of PAAMS*, volume 1233, pages 205–217. Springer.
- Brännström, A., Kampik, T., and Nieves, J. C. (2020). Towards human-aware epistemic planning for promoting behavior-change. In *Workshop on Epistemic Planning (EpiP)@ ICAPS, Online, October 26-30, 2020*.
- Broekens, J., Harbers, M., Hindriks, K., van den Bosch, K., Jonker, C., and Meyer, J.-J. (2010). Do you get it? user-evaluated explainable bdi agents. In Dix, J. and Witteveen, C., editors, *Multiagent System Technologies*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Caminada, M. (2017). Rationality postulates: applying argumentation theory for non-monotonic reasoning. *Journal of Applied Logics*, 4(8):2707–2734.
- Cardoso, R. C. and Ferrando, A. (2021). A review of agent-based programming for multi-agent systems. *Computers*, 10(2):16.
- Cerutti, F., Vallati, M., and Giacomin, M. (2016a). Efficient and off-the-shelf solver: jargsemsat. In *Computational Models of Argument - Proceedings of COMMA*, volume 287, pages 465–466. IOS Press.
- Cerutti, F., Vallati, M., and Giacomin, M. (2016b). Where are we now? state of the art and future trends of solvers for hard argumentation problems. In *Computational Models of Argument - Proceedings of COMMA*, volume 287, pages 207–218. IOS Press.
- Chakraborti, T. (2018). *Foundations of Human-Aware Planning-A Tale of Three Models*. PhD thesis, Arizona State University.
- Chakraborti, T., Sreedharan, S., and Kambhampati, S. (2018). Human-aware planning revisited: A tale of three models. In *IJCAI-ECAI XAI/ICAPS XAIP Workshops*.
- Ciocarlan, A., Masthoff, J., and Oren, N. (2019). Actual persuasiveness: Impact of personality, age and gender on message type susceptibility. In *Persuasive Technology: Development of Persuasive and Behavior Change Support Systems - 14th International Conference, PERSUASIVE 2019, Limassol, Cyprus, April 9-11, 2019, Proceedings*, volume 11433, pages 283–294. Springer.



- Cirillo, M., Karlsson, L., and Saffiotti, A. (2010). Human-aware task planning: An application to mobile robots. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1(2):1–26.
- Cocarascu, O., Rago, A., and Toni, F. (2019). Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In Elkind, E., Veloso, M., Agmon, N., and Taylor, M. E., editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 1261–1269. International Foundation for Autonomous Agents and Multiagent Systems.
- Craandijk, D. and Bex, F. (2020). Deep learning for abstract argumentation semantics. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1667–1673. ijcai.org.
- Čyras, K., Rago, A., Albini, E., Baroni, P., and Toni, F. (2021). Argumentative XAI: A Survey. In Zhou, Z.-H., editor, *30th International Joint Conference on Artificial Intelligence*, pages 4392–4399, Montreal. IJCAI.
- Dalibón, S. F., Martínez, D., and Simari, G. (2012). An approach to emotion-based abstract argumentative reasoning. In *13th Argentine Symposium on Artificial Intelligence*.
- Davis, C., Patte, K., Tweed, S., and Curtis, C. (2007). Personality traits associated with decision-making deficits. *Personality and Individual Differences*, 42(2):279–290.
- de Jonge, D. and Sierra, C. (2017). D-brane: a diplomacy playing agent for automated negotiations research. *Appl. Intell.*, 47(1):158–177.
- Dias, J., Mascarenhas, S., and Paiva, A. (2014). *FAtiMA Modular: Towards an Agent Architecture with a Generic Appraisal Framework*, pages 44–56. Springer International Publishing, Cham.
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358.
- Floyd, M. W., Roberts, M., and Aha, D. W. (2018). Hybrid goal selection and planning in a goal reasoning agent using partially specified preferences. In *Proceedings of the Workshop on Goal Reasoning (held at the 27th Int'l Joint Conf. on AI and the 23rd European Conf. on AI)*.
- Frith, C. and Frith, U. (2005). Theory of mind. *Current biology*, 15(17):R644–R645.
- García, A. J. and Simari, G. R. (2004). Defeasible logic programming: An argumentative approach. *Theory Pract. Log. Program.*, 4(1-2):95–138.
- Gebhard, P. (2005). Alma: a layered model of affect. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 29–36.
- Gelfond, M. and Lifschitz, V. (1998). Action languages.
- Gintis, H. (2018). *Rational Choice Explained and Defended*, pages 95–114. Springer International Publishing, Cham.
- Grecucci, A., Giorgetta, C., Lorandini, S., Sanfey, A. G., and Bonini, N. (2020). Changing decisions by changing emotions: Behavioral and physiological evidence of two emotion regulation strategies. *Journal of Neuroscience, Psychology, and Economics*, 13(3):178.
- Helmert, M. and Geffner, H. (2008). Unifying the causal graph and additive heuristics. In *ICAPS*, pages 140–147.
- Heras, S., Palanca, J., Rodriguez, P., Duque-Méndez, N., and Julian, V. (2020). Recommending learning objects with arguments and explanations. *Applied Sciences*, 10(10):3341.
- Kahneman, D. and Tversky, A. (1979). *Prospect Theory: An Analysis of Decision Under Risk*, chapter Chapter 6, pages 99–127.
- Kampik, T. and Gabbay, D. (2021). Explainable reasoning in face of contradictions: From humans to machines. In Calvaresi, D., Najjar, A., Winikoff, M., and Främling, K., editors, *Explainable and Transparent AI and Multi-Agent Systems*, pages 280–295, Cham. Springer International Publishing.
- Kampik, T., Nieves, J. C., and Lindgren, H. (2019a). Explaining sympathetic actions of rational agents. In Calvaresi, D., Najjar, A., Schumacher, M., and Främling, K., editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 59–76, Cham. Springer International Publishing.
- Kampik, T., Nieves, J. C., and Lindgren, H. (2019b). Implementing argumentation-enabled empathic agents. In Slavkovik, M., editor, *Multi-Agent Systems*, pages 140–155, Cham. Springer International Publishing.
- Köckemann, U., Pecora, F., and Karlsson, L. (2014). Grandpa hates robots-interaction constraints for planning in inhabited environments. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.
- Kökciyan, N., Yaglikci, N., and Yolum, P. (2017). An argumentation approach for resolving privacy disputes in online social networks. *ACM Trans. Internet Techn.*, 17(3):27:1–27:22.
- Kravari, K. and Bassiliades, N. (2015). A survey of agent platforms. *Journal of Artificial Societies and Social Simulation*, 18(1):11.
- Logan, B. (2018). An agent programming manifesto. *International Journal of Agent-Oriented Software Engineering*, 6(2):187–210.
- Mascardi, V., Weyns, D., Ricci, A., Earle, C. B., Casals, A., Challenger, M., Chopra, A., Ciorrea, A., Dennis, L. A., Díaz, A. F., El Fallah-Seghrouchni, A., Ferrando, A., Fredlund, L.-r., Giunchiglia, E., Guessoum, Z., Günay, A., Hindriks, K., Iglesias, C. A., Logan, B., Kampik, T., Kardas, G., Koeman, V. J., Larsen, J. B., Mayer, S., Méndez, T., Nieves, J. C., Seidita, V., Teze, B. T., Varga, L. Z., and Winikoff, M. (2019). Engineering multi-agent systems: State of affairs and the road ahead. *SIGSOFT Softw. Eng. Notes*, 44(1):18–28.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

- Mualla, Y., Tchappi, I., Kampik, T., Najjar, A., Calvaresi, D., Abbas-Turki, A., Galland, S., and Nicolle, C. (2021). The quest of parsimonious xai: a human-agent architecture for explanation formulation. *Artificial Intelligence*, page 103573.
- Ojha, S. and Williams, M.-A. (2017). Emotional appraisal: A computational perspective. In *Fifth annual conference on advances in cognitive systems*. ACS.
- Osborne, M. J. and Rubinstein, A. (2020). *Models in Microeconomic Theory*. Open Book Publishers.
- Paiva, A., Leite, I., Boukricha, H., and Wachsmuth, I. (2017). Empathy in virtual agents and robots: a survey. *ACM Transactions on Interactive Intelligent Systems (TiIS)*, 7(3):1–40.
- Picard, R. W. (1997). *Affective Computing*. The MIT Press.
- Pollock, J. L. (1987). Defeasible reasoning. *Cogn. Sci.*, 11(4):481–518.
- Rao, A. S. (1996). Agentspeak(1): Bdi agents speak out in a logical computable language. In Van de Velde, W. and Perram, J. W., editors, *Agents Breaking Away*, pages 42–55, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Ruiz-Dolz, R. (2020). Towards an artificial argumentation system. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 5206–5207. ijcai.org.
- Ruiz-Dolz, R., Alemany, J., Heras, S., and García-Fornes, A. (2021). On the prevention of privacy threats: How can we persuade our social network users? *CoRR*, abs/2104.10004.
- Ruiz-Dolz, R., Heras, S., Alemany, J., and García-Fornes, A. (2019). Towards an argumentation system for assisting users with privacy management in online social networks. In *Proceedings of the 19th Workshop on Computational Models of Natural Argument, CMNA@PERSUASIVE*, volume 2346, pages 17–28. CEUR-WS.org.
- Seegerstrom, S. C. and Smith, G. T. (2019). Personality and coping: Individual differences in responses to emotion. *Annual review of psychology*, 70:651–671.
- Seyfarth, R. M. and Cheney, D. L. (2015). Social cognition. *Animal Behaviour*, 103:191–202.
- Talamadupula, K., Benton, J., Kambhampati, S., Schermerhorn, P., and Scheutz, M. (2010). Planning for human-robot teaming in open worlds. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1(2):1–24.
- Taverner, J., Vivancos, E., and Botti, V. (2021). A fuzzy appraisal model for affective agents adapted to cultural environments using the pleasure and arousal dimensions. *Information Sciences*, 546:74–86.
- Thimm, M. (2012). A probabilistic semantics for abstract argumentation. In *ECAI 2012 - 20th European Conference on Artificial Intelligence*, volume 242, pages 750–755. IOS Press.
- Thomas, R. J., Masthoff, J., and Oren, N. (2017). Adapting healthy eating messages to personality. In *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors - 12th International Conference*, volume 10171, pages 119–132. Springer.
- Thomas, R. J., Masthoff, J., and Oren, N. (2019). Is argumessage effective? A critical evaluation of the persuasive message generation system. In *Persuasive Technology: Development of Persuasive and Behavior Change Support Systems - 14th International Conference*, volume 11433, pages 87–99. Springer.
- Tulli, S., Correia, F., Mascarenhas, S., Gomes, S., Melo, F. S., and Paiva, A. (2019). Effects of agents’ transparency on teamwork. In Calvaresi, D., Najjar, A., Schumacher, M., and Främling, K., editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 22–37, Cham. Springer International Publishing.
- Čyras, K., Badrinath, R., Mohalik, S. K., Mujumdar, A., Nikou, A., Previti, A., Sundararajan, V., and Feljan, A. V. (2020). Machine reasoning explainability. *arXiv preprint arXiv:2009.00418*.
- Wooldridge, M. (1997). Agent-based software engineering. *IEE Proceedings-Software Engineering*, 144(1):26–37.
- Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H. H., and Kambhampati, S. (2017). Plan explicability and predictability for robot task planning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1313–1320.