Towards Reflective Normative Agents

Nathan Lloyd[0000−0002−7127−2500] and Peter R. Lewis[0000−0003−4271−8611]

Trustworthy AI Lab, Ontario Tech University, Oshawa, Canada
{fname.lname}@ontariotechu.ca

Abstract. An 'Interdisciplinary Frontier' of norm research has emerged within the social sciences, carrying forward a wave of new theories and evidence that describe the conditions required for norms to be measured, represented, spread, and changed. These have informed requirements for a normative agent with mental representations, able to reason about models of self, others, environment, and society. Recent Socio-Cognitive theories of reflection have highlighted the need for high-level reasoning processes to assess whether actions are congruent with prevailing norms, to evaluate beliefs, and to adjust behavior accordingly. Agents lacking these capacities merely engage in passive norm following or compliance. In contrast, we propose a framework encompassing the cognitive abilities necessary for learning, reasoning, and reflecting upon norms beyond mere adherence or compliance. Furthermore, we discuss the necessity for normative agents to be situated in complex, open-ended scenarios from which rich social interactions can emerge autonomously.

Keywords: Reflection · Norms · Cognitive agents.

1 Introduction

Artificial Intelligence has been defined as making "computers do the sorts of things that minds can do" [8, p.1]. Boden's definition evokes questions of whether mind-like faculties can be replicated and to what extent they may modify individual and collective behavior. This interplay between the mind's architecture and behavior observation requires an interdisciplinary approach to developing intelligent agents. This paper discusses the concepts of normative, self-organizing, and reflective agents to propose a framework for capturing the cognitive abilities required to learn, reason, and reflect on norms, not simply follow or comply. The objective is to showcase that reflective normative agents operating in dynamic and complex scenarios can foster the emergence of rich social interactions autonomously, beyond those which do under the assumption that agents are simply maximizing competence or compliance.

Norms can be defined as "the informal rules we live by" [7], ubiquitous within society, and argued as powerful constructs for influencing behavior [23], and also discussed as informal institutional rule types [31, p.14]. Institutions are systems for organizing and standardizing behavior; their structured rules regulate social behavior and have long been recognized as essential mechanisms for collective action, even when individuals do not share a common purpose [30]. Ostrom's [29] pioneering research in shared common-pool resources discussed sustained institutions and cooperative behaviors through self-organization. Self-organization

proposes that individuals create, employ, and modify their institutions to achieve sustainable cooperation [32], mirroring the human potential to change the rules of social interaction [32]. Strides within the social sciences have pushed forward a new constructivist theory of norms [5]. Bicchieri has identified an 'Interdisciplinary Frontier' of norm research [7], compiling theories and empirical science to discuss the significance of norms within society and discern why individuals may choose to follow social norms. Although this new wave of research provides a map towards the development of normative agents through its formalizations, critical mind-like qualities integral for normative agents are yet to be incorporated.

In discussing the issues of incomplete minds, Lewis and Sarkadi [24] call attention to the failings of many modern AI systems and their inability to reflect upon the social and ethical nature of their decisions. Reflection is a core mental mechanism that motivates the evaluation of beliefs, values, and behavior, essential to assess whether one is congruent with prevailing norms [24] and to reason about the mental states of others. For sustainable self-organization and self-governance, agents require the capacity for reflection [37]. However, it is also essential to understand how norms guide behavior, to formalize how they are represented, learned, activated, and updated [26]. This paper builds upon Bicchieiri's formalized constructivist theory of norms [5], a diverse range of intelligent agent research, and the work of Lewis and Sarkadi [24] to define a reflective normative agent architecture for simulating self-organizing behavior.

2 Norms

First, we introduce Bicchieri's formalization of norms, the conditions to support them, and the necessary properties for norm compliance.

2.1 Components of Norms

Despite Bicchieri's assertion that there is a lack of agreement regarding the influence of norms [5, p.1], she has done much to clarify the once obscure and muddied definitions that orbit the discussion around norms [5, 6], replacing ambiguity with concrete formalizations. In The Grammar of Society [5, p.2], Bicchieri presents a 'constructivist' theory that defines norms in terms of *expectations* and *preferences*. Expectations and preferences are the building blocks of many social constructs; as such, they can be considered integral components for designing and developing artificial social systems. We will first introduce a *personal normative belief*, a concept represented through deontic sentences, and describes what "I believe (I/We) ought to do..." *Social expectations* is an umbrella term encompassing two different types of expectations. The first, *empirical expectations*, is defined as a belief about another's future behavior based on past behavior [6, p16-17], written in the form "I expect they're going to..." The second, *normative expectations*, can be described as a second-order belief about another's personal normative belief, commonly expressed through deontic sentences such as "I believe that most people think we ought to do..." *Preferences* refer to an

individual's disposition to behave in a certain way within a specific context, indicating how expectations alone may not necessarily impact behavior. Preferences may be described as socially unconditional, where others do not influence one's choice, or as conditional, by dependence upon empirical and normative expectations. It is common to see an additional distinction made in Bicchieri's work that an individual's preferences and expectations are bound to a reference network. A *reference network* is a set of individuals who matter to our decision-making processes, highlighting the interdependency of behaviors.

Descriptive Norms Using Bicchieri's building blocks, a descriptive norm is defined as a pattern of behavior that an individual prefers to engage in on the condition that others within their reference network also engage in it [5]. A descriptive norm describes interdependent behaviors where preferences are conditional upon empirical expectations alone. Following this distinction, descriptive norms drive behaviors such as imitation and coordination, as they are based solely on the behavior of others and not on another's normative expectations.

Social Norms Whereas descriptive norms are composed of empirical expectations and conditional preferences alone, social norms require the addition of normative expectations. Social norms are interdependent, socially conditional, rely upon social expectations, and require that individuals acknowledge the existence of the normative rules and to which situation they should be applied. Bicchieri [5, p.11] defines the conditions for a social norm to exist as follows:

> Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. We say that R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:
> *Contingency*: i knows that a rule R exists and applies to situations of type S;
> *Conditional Preference*: i prefers to conform to R in situations of type S on the condition that:
> *(a) Empirical Expectations*: i believes that a sufficiently large subset of P conforms to R in situations of type S;
> and either
> *(b) Normative Expectations*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
> or
> *(b') Normative Expectations with Sanctions*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform, and may sanction behavior.

It is essential to disambiguate the notation of P, which, dependent upon simulation objectives, may represent a reference network or a larger population. This distinction is crucial when applying *behavioral rules*, as one rule may be a social norm in P and not in P'. Bicchieri outlines the conditions for a social

norm to exist by requiring a sufficiently large subset of *conditional followers* P_{cf}, a conditional follower, however, merely recognizes the existence of the norm and is said to become a *follower* when their social expectations are fulfilled. We can then say the norm is *followed* if a sufficiently large subset P_f of P_{cf} meets the conditions of contingency, conditional preference, and social expectations: $P_f \subseteq P_{cf} \subseteq P$.

Not Norms Finally, in light of often ambiguous and misrepresented terms surrounding the discussion of norms, it is essential to discuss apparently similar concepts. Outlining Bicchieri's building blocks reveals the factor distinguishing normative from non-normative behaviors like customs, habits, shared morals, and religious rules. This factor is interdependency. It is a common pitfall to erroneously group normative and non-normative behaviors together. However, it is essential to note that independent and interdependent behaviors are motivated by entirely different sets of preferences, with independent actions occurring irrespective of what others do.

2.2 Requirements for Norm Competence

Inspired by Bicchieri's definition of norms, recent work by Malle et al. [26] discuss the properties required for an artificial agent with norm competency; here we discuss their propositions in light of Bicchieri's formalizations.

Norm Representation The language utilized in discussing normative expectations connotes the use of deontic logic. It is common to express these statements as what one should or ought to do or what is obligatory, optional, permissible, or prohibited. Malle et al. [26] propose that normative rules be represented through three distinct categories: prescription, prohibition, and permissions. They suggest a graded ordinal scale to provide granular insight into the demand of the normative rules, which signals the strength of the expectation. Grading normative rules provides an intuitive mechanism for decision-making. The scale provides a weight such that an agent can distinguish to what degree the norms are demanded, i.e., recognizing whether prescriptions are required or suggested. Norms are seldom described as absolutes, their supporting language often possessing a fuzzy and qualitative nature; thus, graded demand becomes appropriate.

Context Sensitivity Bicchieri's *contingency* condition requires an agent to be aware of a *behavioral rule* and to which *situations* they are applied for their activation. Recognizing a situation implies that an agent must first be able to perceive its environment and then infer what features within the context activate the norm; situational cues. Situational cues may come from the environment and others within that situation, activating one's beliefs, preferences, and any known accompanying norms. The precise mechanism by which humans perceive contexts and how such contexts trigger the applicable norms is presently unclear [26]; however, this is expected to be computationally demanding for non-humans [35].

Prevalence A requirement for identifying social norms is to recognize their prevalence within a reference network. Although a definition for social norm followers provides an understanding of the prevalence of a norm, this knowledge cannot be assumed to be available to individual agents. Therefore, prevalence can be calculated or estimated based on what is observed or communicated.

Norm Learning and Updating Norm learning illustrates the cyclical relationship between external and internal norms [11], where external events influence one's internal representation, and in turn, the internal norm shapes one's behavior. For example, external information may come from explicit instructions, such as signs, verbally communicated rules, or expressing (dis)approval when a norm is either conformed or transgressed. These communications may infer the rule's demand, with stronger sanctions and continued communication of that rule highlighting its significance to the reference network. Observing others' behavior and the consequences of their actions provides another vector from which to learn, but this may be insufficient for learning norms accurately. For example, observing behaviors does not express the individual's desire or motivation; they may be self-interested and act independently of others or be compliant due to pluralistic ignorance. Observations are also limited in realistic situations where agents can only access imperfect information from their surroundings. Thus, to not confuse norm-guided behavior and an individual's goals or desires, one can learn from the consequences of actions, whether a behavior is reinforced or sanctioned. However, the enforcement of social norms can vary significantly [5, p.8] and may be heavily influenced by the interdependency of the reference network [21]. Norm learning thereby describes an observational process to update one's own mental representations and beliefs about others.

3 Towards Reflective Normative Agents

The concepts presented thus far facilitate an agent's ability to acquire, represent, and adhere to norms. However, these concepts primarily ensure competency or compliance, leaving no room for an agent to intentionally violate a norm, which may be advantageous and preferable for achieving an individual's or society's goals [12, 10]. Before Bicchieri's formalizations, Castelfranchi et al. [12] discussed the necessity of *intelligent violations* and the requirement for cognitive agents that may form mental representations of beliefs, goals, and intentions, an aspect missing in recent prior work. Reasoning about these mental representations requires further discussion about different reasoning processes. Bicchieri's work highlights the tendency to focus on deliberation, and higher-level reasoning capabilities like reflection appear absent in the discussion of normative agents.

Open-ended environments entail scenarios where agents may need to learn to coordinate, cooperate, conform, or control one another [3], learning appropriate strategies and self-organizing through their interactions with the environment and one another. Moreover, open-ended situations do away with domain constraints and the specification of narrow problems that may otherwise constrain

norms and emergent group behavior. Open-ended, complex environments have seen particular success in developing deep learning agents [28, 34]. Agent-based modeling is a "quintessential tool for open-ended social theorizing" [13], where outcomes (like norms) are socially constructed, emerging organically from social interaction. Indeed, norms are emergent phenomena that come in all shapes and sizes, varying tremendously worldwide due to the complexity of the environment upon which human societies sit. Therefore, it is necessary to model the conditions that initially allow various norms to arise. We propose open-ended situations in which there is no single task. Instead, the emergent norms and behaviors are determined by the initialization of the world and its inhabiting agents.

3.1 Theory of Mind

Social expectations are beliefs about others' behavior and what others believe. To reason about these, an agent must possess models of others that appropriately incorporate their mental states to reason about the prevalence of norms. The cognitive science community has extensively investigated the process of forming mental representations of the goals, beliefs, and preferences of those who are interacted with [40, 15]. The capability to construct mental models of others, known as the Theory of Mind (ToM) [40], is a fundamental aspect of human social intelligence [36]. ToM has multiple orders [18], but social expectations require the second. Zeroth order ToM states that an individual can reason about their knowledge, beliefs, desires, and perceived state of the world but maintain no understanding of the mental state of others [18]. First-order ToM involves recognizing that one and others have desires and beliefs that influence behavior. Second-order ToM recognizes that others may hold beliefs about oneself; essential for normative expectations. The ability to infer the intentions and form beliefs about other agents from observable actions has significant practical applications, particularly for normative agents in cooperative and competitive tasks [27].

Modeling others also provides additional qualities that describe how social norms are adhered to. *Social-image* and *Self-image* are aspects that are congruent with the ToM. These mechanisms facilitate functionalities that help to describe adherence to a social norm in private and public settings [17, 7]. A social image concern refers to an individual's desire to appear in a particular light to others within their reference network and to seek their approval [38]. A social image enables individuals to be aware of how they or others are perceived, determining who may be trustworthy, reliable, and reputable, factors influencing human decision-making [2]. Self-image is discussed as a mechanism to explain why individuals may continue to exhibit normative behaviors in a private setting. Rather than exhibiting idealistic characteristics and behaviors for the benefit of others, individuals choose to adhere to norms to reinforce a positive self-image; to feel good about themselves. Individuals' self and social image concerns infer self-awareness about their behavior and how others perceive them within their reference network. Without explicit coordination protocols, modeling other agents becomes an essential skill for effective collaboration [2] and enables the achievement of common goals with decreased effort [36].

3.2 Diverse Reasoning Capabilities

The social phenomena and properties of norms discussed thus far imply the requirement for a cognitive agent, a type of agent that can emulate the human capacity for memory and problem-solving. This distinction connotes a *stronger* notion of autonomy for agents [41], those characterized by *mentalistic* concepts, able to manipulate and reason upon mental presentations like goals, beliefs, and context. This contrasts the implementation suggested by Malle et al. [26], who define a norm conflict resolution property to obey as many norms as possible. Instead, a cognitive agent would reason about their mental representations to decide whether they should conform or transgress. An agent relies upon its reasoning processes to make these decisions, with many architectures proposed for domain-specific problems [4]. There have been many implementations of cognitive agents, but they are argued to be divided into two overarching approaches [24], explicit architectures and emerging cognition from complex systems. The Belief-Desire-Intention Architecture is a commonly used example for explicitly defined cognitive agents [33], an architecture for mental representations that support cognitive reasoning. Systems built using Artificial Neural Networks have also been successful in developing cognitive agents [39]; similar to humans, it is expected that cognition may emerge through model complexity. However, there are concerns that deep learning approaches may develop shortcuts that impede accurate mental representations in ToM research [3].

The standard view for information processing and decision-making is the deliberative route to behavior, a conscious process that weighs each factor against an individual's preferences to determine an outcome. There has been much work in this regard explicitly for normative agents [12, 25], where deliberation is used to consciously reason and decide whether to conform or transgress rather than norm following through some hard-coded filter or goal of maximization. In humans, this process is costly, requiring time, skill, and effort to systematically weigh all factors and calculate the potential utility of available strategies. As such, the deliberative route to behavior has received criticism for being an overcited but underused decision-making method [5, p.4-7]. The over-simplistic view that individuals weigh all their decisions and outcomes is unjustified when considering how some decisions are made instinctively through some reactive or unconscious process. This mode of thinking is dubbed the heuristic route. This information processing method calls upon an in-memory set of rules to prescribe actions based on perceived contextual information, beliefs, desires, and expectations. The heuristic method is strengthened by cognitive shortcuts, where overlapping contextual cues and classes of similar situations allow individuals to generalize or extrapolate one behavioral rule to a new situation. These two modes of information processing are well discussed within the literature, often described as thinking fast and slow [22]; both are thought to be simultaneously occurring in some form or another.

Despite both processes being intuitive for information processing and decision-making procedures, Bicchieri argues that they are incompatible given that the former considers preferences as mental states and the heuristic approach does

not [5, p.6]. Between these two modes lies the dispositional approach, a philosophical tradition that considers beliefs and desires in appropriate circumstances. A dispositional decision-making process infers that individuals will be motivated to act according to their preferences until they are dissatisfied by the outcome of their actions or others, sparking a reflective process. A reflective process is essential for dealing with ambiguity, emergent knowledge, and social context [24]. The dispositional process with reflection reveals how a default behavior (heuristic) may be followed until an individual feels unfulfilled based on their expectations, invoking a conscious and reflective process to adjust rules, beliefs, and goals.

3.3 Reflection

Beyond the requirements for a normative agent, we posit the necessity for a reflective capability within normative agents. Reflection is a higher-level reasoning process than previously discussed, enabling individuals to deliberate on abstract concepts like beliefs, behaviors, and norms concerning actions taken and their outcomes. A reflective agent reasons about their behavior [9], the behavior of others, and the external world. A reflective process is essential for determining whether one's actions were congruent with prevailing norms [24], for dealing with ambiguity, emergent knowledge, and reasoning about social contexts. Lewis and Sarkadi introduce a novel socio-cognitive theory of reflection in artificial intelligence [24], which outlines different tiers of reflective capabilities and the corresponding qualities necessary to attain each tier. Expanding on Hesslow's Simulation Theory of Cognition [19, 20], Lewis and Sarkadi explore the role of simulation and hypothesis testing as reflective processes, which follows seminal cognitive science research that discusses an individuals ability to simulate the behavior of others by adopting their perspective [16, 14], enabling them to comprehend the intentions or motives of others and respond appropriately in social contexts. In discussing Ostrom's work and the Tragedy of the Commons, Powers et al. [32] call attention to the capability of reflective processes within humans, highlighting how this mechanism enables individuals to "change the rules of the game" and avoid undesirable outcomes; recent work highlights the success reflective self-governance for sustainability [1]. Reflection is a crucial cognitive component for normative agents, essential for reasoning about social expectations, normative rules, models of self, others, the environment, and society, correcting wrong beliefs, and motivating new goals and behaviors.

4 Agent Specification

Situated agents can possess the fundamental capacities of perception, locomotion, and interaction. Normative agents extend this ability to ***observe*** the environment and the actions of those within it, ***form beliefs*** about the behavior of others, ***recognize*** the existence of normative rules and to which ***context*** they are applied, hold a model for ***reference networks***, and ***evaluate*** their adherence to a norm based on achieving their own ***goals***, for example, maintaining

satiation. Furthermore, normative agents require the capacity to communicate (directly or through signals) and interpret information from the environment and one another [36]. The open-ended and undefined configurable environments an agent may inhabit make defining an agent's full requirements challenging, with conditional agent requirements dependent on the situation. However, we can consider the agent's and environment's basic properties to motivate emergent behaviors like cooperation, coordination, conformity, and control. Beyond capabilities associated with norms and being situated, a reflective normative agent can ***reflect*** upon its mental representations and update or formulate new normative rules, beliefs, and goals. Reflection can be considered an intentional action triggered during a *shock* following a negative outcome when facing ambiguity in a dispositional reasoning process or an action taken during times of comfort. Given these conditions, we define a normative system as follows. Let $M = (G, E)$ be a formal model where E is the environment, and G is the global population of agents. We can then define the composition of an individual agent $i \in G$ as:

$$(S_i, B_i, D_i, X_i)$$

- S_i is the set of i's observations of their own and other's behavior and state;
- B_i is the set of i's beliefs;
- D_i is the set of i's goals or desires;
- X_i is the set of actions known to i (capabilities);

This notation provides an intuitive and generalizable abstraction for normative agents, with i's knowledge, beliefs, goals, and abilities different from those of another agent j. This formalization maintains a level of abstraction for X_i and D_i, which will remain undefined for open-ended scenarios and open to relevant instantiation for a given situation. It can then be stated that an individual's observations denoted as S_i would inform their beliefs B_i. Given the requirements outlined prior, agent i's beliefs can be stated as follows:

$$B_i = (C_i, P_i, R_i, O_i, E_i, N_i, Q_i, A_i, W_i)$$

- P_i is the set of reference networks known to i, the groups to whom their decisions matter;
- R_i is the set of behavioral rules known by i and to which P they belong;
- O_i is the set of models of agents known to i;
- E_i is the set of empirical expectations i has regarding S_i, R_i, and P_i;
- N_i is the set of normative expectations i has regarding S_i, R_i, and P_i;
- Q_i is the set of personal normative beliefs i has regarding S_i, R_i, and P_i;
- A_i is i's model of self;
- W_i is i's model of the world;

Congruent with Bicchieri's formal model of social norms, an agent's belief model incorporates the aforementioned building blocks for the many social constructs to exist. Beyond these requirements, an agent contains a model of itself A_i, the world W_i, and others O_i to facilitate the cognitive requirements and mechanisms

attributed to why individuals may choose to conform or transgress. An agent's model of self may contain characteristics such as risk sensitivity, self-efficacy, or their tendency to seek approval, as well as their self-image and social image. An agent's world model reflects their incomplete knowledge of the state of the world through their perception. Our requirements for a normative agent imply the need to model others, necessitating predictions about another's intentions or goals [5, p.56]. As such, a model of others, O_i, is the final layer to unpack. Where i's beliefs about another agent j is stated as:

$$O_i^j = (P_i^j, D_i^j, X_i^j, O_i^{O_j}, R_i^j, Q_i^j, A_i^j)$$

- P_i^j is i's beliefs of j's membership to reference networks, the groups to whom i believes impacts j's decisions;
- D_i^j is i's beliefs of j's goals or desires;
- X_i^j is i's beliefs of actions known to j's (capabilities);
- $O_i^{O_j}$ is i's beliefs about j's model of others;
- R_i^j is the set of behavioral rules i believes j to know;
- Q_i^j is i's beliefs of j's personal normative beliefs;
- A_i^j is i's beliefs of how j perceives i;

This specification does not explicitly capture the processes for learning, updating, and reflection; these will be explored in future work via operationalization.

5 Conclusion

Our exploration into various aspects of reflective normative agents has shed light on the complexity and importance of understanding human behavior and decision-making processes. Through extending Bicchieri's formalizations, we have delved into the essential components that a normative agent can possess. These include incorporating mental representations of self, others, environment, and society, considering multiple modes of reasoning, and the significant role of reflection in shaping an agent's actions. Furthermore, we have underscored the importance of creating dynamic and complex environments that are open-ended, allowing for the emergence of rich self-organizing behavior. To appreciate the complexity of reflective normative agents, it will be necessary to situate agents in conditions that motivate social constructionism. This highlights the need for a unified testbed containing sufficiently complex and open-ended scenarios to study agent capabilities. Through our exploration, we have come to appreciate the multifaceted nature of normative and cognitive agents, recognizing the complexity of human behavior and the need for nuanced modeling. By addressing these requirements and leveraging the power of reflection, we can develop more advanced agents that demonstrate autonomous decision-making and exhibit the richness of social constructionism.

References

1. Aishwaryaprajna, Lewis, P.R.: Exploring intervention in co-evolving deliberative neuro-evolution with reflective governance for the sustainable foraging problem. In: Artificial Life (2023), to Appear

2. Albrecht, S.V., Stone, P.: Autonomous agents modelling other agents: A comprehensive survey and open problems. Artificial Intelligence **258**, 66–95 (2018)
3. Aru, J., Labash, A., Corcoll, O., Vicente, R.: Mind the gap: challenges of deep learning approaches to Theory of Mind. Artificial Intelligence Review (2023)
4. Balke, T., Gilbert, N.: How do agents make decisions? a survey. Journal of Artificial Societies and Social Simulation **17**(4), 13 (2014)
5. Bicchieri, C.: The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press (2005)
6. Bicchieri, C.: Norms in the Wild: How to Diagnose, Measure, and Change Social Norms. Oxford University Press (2017)
7. Bicchieri, C., Dimant, E., Gelfand, M., Sonderegger, S.: Social norms and behavior change: The interdisciplinary research frontier. Journal of Economic Behavior & Organization **205** (2023)
8. Boden, M.A.: AI: Its Nature and Future. Oxford University Press (2016)
9. Brazier, F., Wijngaards, N.: Designing self-modifying agents. In: Gero, J. (ed.) Computational and Cognitive Models of Creative Design V. pp. 93–112. Key Centre of Design Computing and Cognition, University of Sydney (2001)
10. Burth Kurka, D., Pitt, J., Lewis, P.R., Patelli, A., Ekárt, A.: Disobedience as a mechanism of change. In: 2018 IEEE 12th International Conference on Self-Adaptive and Self-Organizing Systems (SASO). pp. 1–10 (2018)
11. Castelfranchi, C.: A cognitive framing for norm change. In: Dignum, V., Noriega, P., Sensoy, M., Sichman, J.S. (eds.) Coordination, Organizations, Institutions, and Norms in Agent Systems XI. pp. 22–41. Springer International Publishing, Cham (2016)
12. Castelfranchi, C., Dignum, F., Jonker, C.M., Treur, J.: Deliberative Normative Agents: Principles and Architecture. In: Goos, G., Hartmanis, J., van Leeuwen, J., Jennings, N.R., Lespérance, Y. (eds.) Intelligent Agents VI. Agent Theories, Architectures, and Languages, vol. 1757, pp. 364–378. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)
13. Devereaux, A., Wagner, R.E.: Agent-based modeling as quintessential tool for open-ended social theorizing. Tech. Rep. 19-07, George Mason University, Fairfax, VA (2019)
14. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences **2**(12), 493–501 (1998)
15. Gopnik, A., Wellman, H.M.: Why the Child's Theory of Mind Really Is a Theory. Mind & Language **7**(1-2), 145–171 (1992)
16. Gordon, R.M.: Folk psychology as simulation. Mind & Language **1**(2), 158–171 (1986)
17. Gross, J., Vostroknutov, A.: Why do people follow social norms? Current Opinion in Psychology **44**, 1–6 (2022)
18. Hedden, T., Zhang, J.: What do you think i think you think?: Strategic reasoning in matrix games. Cognition **85**(1), 1–36 (2002)
19. Hesslow, G.: Conscious thought as simulation of behaviour and perception. Trends in Cognitive Sciences **6**(6), 242–247 (2002)
20. Hesslow, G.: The current status of the simulation theory of cognition. Brain Research **1428**, 71–79 (2012)
21. Horne, C.: Explaining Norm Enforcement. Rationality and Society **19**(2), 139–170 (2007)
22. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York, NY (2011)

23. Kalkstein, D.A., Hook, C.J., Hard, B.M., Walton, G.M.: Social norms govern what behaviors come to mind-And what do not. Journal of Personality and Social Psychology (2022)
24. Lewis, P.R., Sarkadi, S.: Reflective artificial intelligence (2023), https://arxiv.org/abs/2301.10823
25. y López, F.L., Luck, M., d'Inverno, M.: A normative framework for agent-based systems. Computational and Mathematical Organization Theory **12**(2-3), 227–250 (2006)
26. Malle, B.F., Bello, P., Scheutz, M.: Requirements for an artificial agent with norm competence. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. p. 21–27. AIES '19, Association for Computing Machinery, New York, NY, USA (2019)
27. Matiisen, T., Labash, A., Majoral, D., Aru, J., Vicente, R.: Do deep reinforcement learning agents model intentions? Stats **6**(1), 50–66 (2023)
28. Open Ended Learning Team, Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., McAleese, N., Bradley-Schmieg, N., Wong, N., Porcel, N., Raileanu, R., Hughes-Fitt, S., Dalibard, V., Czarnecki, W.M.: Open-ended learning leads to generally capable agents (2021)
29. Ostrom, E.: Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press (1990)
30. Pitt, J., Diaconescu, A., Bollier, D.: Technology for collective action [special section introduction]. IEEE Technology and Society Magazine **33**(3), 32–34 (2014)
31. Polski, M.M., Ostrom, E.: An institutional framework for policy analysis and design (1999)
32. Powers, S.T., Ekárt, A., Lewis, P.R.: Modelling enduring institutions: The complementarity of evolutionary and agent-based approaches. Cognitive Systems Research **52**, 67–81 (2018)
33. Rao, A.S., Georgeff, M.P.: Bdi agents: From theory to practice. In: International Conference on Multiagent Systems (1995)
34. Samvelyan, M., Kirk, R., Kurin, V., Parker-Holder, J., Jiang, M., Hambro, E., Petroni, F., Küttler, H., Grefenstette, E., Rocktäschel, T.: Minihack the planet: A sandbox for open-ended reinforcement learning research (2021)
35. Scheutz, M., Malle, B.: Moral Robots, pp. 363–377. Routledge/Taylor & Francis Group (2017)
36. Sclar, M., Neubig, G., Bisk, Y.: Symmetric machine theory of mind. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 19450–19466. PMLR (2022)
37. Scott, M., Pitt, J.: Interdependent Self-Organizing Mechanisms for Cooperative Survival. Artificial Life pp. 1–37 (2023)
38. te Velde, V.L.: Heterogeneous norms: Social image and social pressure when people disagree. Journal of Economic Behavior & Organization **194**, 319–340 (2022)
39. Volzhenin, K., Changeux, J.P., Dumas, G.: Multilevel development of cognitive abilities in an artificial neural network. Proceedings of the National Academy of Sciences **119**(39), e2201304119 (2022)
40. Woodruff, G., Premack, D.: Intentional communication in the chimpanzee: The development of deception. Cognition **7**(4), 333–362 (1979)
41. Wooldridge, M., Jennings, N.R.: Agent theories, architectures, and languages: A survey. In: Wooldridge, M.J., Jennings, N.R. (eds.) Intelligent Agents. pp. 1–39. Springer Berlin Heidelberg, Berlin, Heidelberg (1995)