I Know I Should: Normative Competence From Biology To AI

Joel Z. Leibo (jzl@deepmind.com)
Google DeepMind
London, UK

Sydney Levine (sydneyl@allenai.org)
Department of Psychology, New York University
New York, NY, USA

John Michael (john.michael@unimi.it)
Department of Philosophy, University of Milan
Milano, Italy

Jordan Theriault (jo.theriault@northeastern.edu)
Department of Psychology, Northeastern University
Boston, MA, USA

Luca Tummolini (luca.tummolini@istc.cnr.it)
Institute of Cognitive Sciences and Technologies, CNR
Roma, Italy

Keywords: social norms, normative competence, sense of obligation, commitments, sense of should, moral psychology.

Overview

How do people understand what is normatively expected from them? And how does normative cognition motivate action? A distinct feature of human sociality is our capacity to sense the *appropriate* thing to say in a certain context or the action that *should* be done and to care enough to do it. Such competence with social norms stabilizes joint action between dyads and enables cooperation among thousands, but whether it is the product of domain-general learning mechanisms or dedicated cognitive ones is still unclear (Heyes, 2024). Moreover, as the many blunders made by contemporary AI systems like LLMs show, autonomous AI agents too need to acquire some level of normative competence to be reliable partners. However whether this is possible with current architectures is still debated (Browning, 2024).

The aim of this symposium is to bring together researchers to discuss the biological, cognitive, computational and interactive bases of normative motivation and judgment. The symposium is organized around four different interdisciplinary perspectives. **Jordan Theriault**, Assistant Professor in Psychology and Biology at Northeastern University, offers a novel and biologically inspired computational model of normative motivation grounded in the predictive view of the brain (Theriault, Young, & Barrett, 2021). **John Michael**, Associate Professor in Philosophy and Psychology at University of Milan, and **Luca Tummolini**, Research Director in cognitive science at the Italian National Research Council, assess the "sense of commitment" that arises in joint action (Michael, Sebanz, & Knoblich, 2016; Andrighetto, Grieco, & Tummolini, 2015) and present experimental evidence and a computational model of its intrinsic motivational force, which may support the formation of social norms (Michael & Tummolini, 2025). **Sydney Levine**, Assistant Professor in the Psychology Department at NYU working at the intersection of cognitive science and AI safety, proposes a new theory of moral cognition in the form of a psychological contractualism that can explain moral norms can get established and adapted (Levine, Chater, Tenenbaum, & Cushman, 2024). Finally, **Joel Z. Leibo**, Senior Research Scientist at Google DeepMind, presents a new theory of the "sense of appropriateness" that can be shared between humans and generative AI agents like LLMs (Leibo et al., 2024).

Symposium Contributions

A dynamic model of social learning and control

Jordan Theriault, Robert Passas, Joshua Paul Rodriguez, Eli Sennesh, and Brennan Klein

Controlling nonlinear dynamic systems is a difficult computational problem with no general solution; yet, humans and other animals control these systems all the time: they control each other. For example, if I asked you to pass the salt at dinner, you probably would. Similarly, abstract social norms can compel you to adopt or avoid behaviors without even being explicitly asked (Theriault et al., 2021). In this talk, we outline a computational model that describes a general strategy that humans and other animals may use to control each other. The answer is not intuitive: to be a good controller, an agent must also open itself to being controlled by others. By temporarily giving up its own goals, and by entering into a predictable equilibrium with a partner, an agent can create the necessary conditions for learning about others and directing their behavior. This strategy may describe the emergence of social norms, and has potential implications for the how machine models might approximate the kind of socially-embedded "general" intelligence that still evades state-of-the-art large language models.

The sense of commitment: from joint action to social norms

John Michael, Luca Tummolini, Giulia Andrighetto, Francesco Mannella, Marcell Szekely, and Julian Zubek

As humans, we are remarkable for the versatility and flexibility with which we coordinate our actions in space and time. Whether in small-scale joint actions – as when two partners cook meals, dance or assemble furniture together – or in large-scale collective actions – as when we organize to mitigate climate change – our affinity for coordination enables us to achieve our goals more efficiently than we otherwise could, and in many cases to achieve goals that

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

could not otherwise be achieved. It also requires us to rely upon others to act as we expect them to, and sometimes to persist in carrying out actions because others are expecting and relying on us to do so. But how do we know when to rely on others? And how do we determine when to persist in carrying out actions that others are relying on us to perform? We present findings from a range of experiments supporting the hypothesis that joint action inherently gives rise to a sense of commitment, which boosts the motivation to persist and thereby stabilizes joint action even when agents' interests are not perfectly aligned and provides the ground for the formation of social norms. We go on to spell out a computational model which illuminates the mechanism whereby joint action generates a sense of commitment: coordination requires joint action partners to adapt their actions according to their expectations about what each other will do, making those expectations manifest and eliciting a motivation to conform to avoid disappointing one's joint action partner.

Resource-rational contractualism: A triple theory of moral cognition

Sydney Levine, Nick Chater, Joshua Tenenbaum, and Fiery Cushman

Our view starts with the assumption that people are self-interested rational actors with conflicting interests who need each other's help to survive and thrive. Morality, then, is motivated by people's (self-interested) attempts to reach agreements of mutual benefit—agreements that will benefit each person, thereby ensuring buy-in to the arrangement. This "contractualist" perspective on the *function* of morality has wide-spread support in philosophy, psychology, economics, biology, and cultural evolution. But how are such agreements defined and understood, on a *proximate* level, as a matter of cognitive mechanism? As a practical matter, after all, investing time and effort in negotiating every interpersonal interaction to come up with an appropriate agreement is unfeasible. Instead, our view, "Resource Rational Contractualism," proposes that people use abstractions and heuristics to efficiently identify mutually beneficial arrangements. We argue that many well-studied elements of our moral minds, such as reasoning about others' utilities ("consequentialist" reasoning) or evaluating intrinsic ethical properties of certain actions ("deontological" reasoning), can be naturally understood as resource-rational approximations of a contractualist ideal. Moreover, this view explains the flexibility of our moral minds—how our moral rules and standards get created, updated and overridden and how we deal with novel cases we have never seen before. Thus, the apparently fragmentary nature of our moral psychology—commonly described in terms of systems in conflict—can be largely unified around the principle of finding mutually beneficial agreements under resource constraint. Our resulting "triple theory" of moral cognition naturally integrates contractualist, consequentialist and deontological concerns.

A theory of appropriateness with applications to generative artificial intelligence

Joel Z. Leibo

What is appropriateness? Humans navigate a multi-scale mosaic of interlocking notions of what is appropriate for different situations. We act one way with our friends, another with our family, and yet another in the office. Likewise for AI, appropriate behavior for a comedy-writing assistant is not the same as appropriate behavior for a customer-service representative. What determines which actions are appropriate in which contexts? And what causes these standards to change over time? Since all judgments of AI appropriateness are ultimately made by humans, we need to understand how appropriateness guides human decision making in order to properly evaluate AI decision making and improve it. This contribution presents a theory of appropriateness: how it functions in human society, how it may be implemented in the brain, and what it means for responsible deployment of generative AI technology.

Acknowledgments

John Michael and Luca Tummolini received funding from the Ministry of University and Research (MUR) and European Union - Next Generation EU [PRIN 2022 PNRR; Prot P2022YYRK3; JM CUP G53D23007310001 and LT CUP B53D23030380001].

References

Andrighetto, G., Grieco, D., & Tummolini, L. (2015). Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. *Frontiers in psychology*, *6*, 1413.

Browning, J. (2024). Getting it right: the limits of fine-tuning large language models. *Ethics and Information Technology*, *26*(2), 1–9.

Heyes, C. (2024). Rethinking norm psychology. *Perspectives on Psychological Science*, *19*(1), 12–38.

Leibo, J. Z., Vezhnevets, A. S., Diaz, M., Agapiou, J. P., Cunningham, W. A., Sunehag, P., ... Osindero, S. (2024). A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*.

Levine, S., Chater, N., Tenenbaum, J. B., & Cushman, F. (2024). Resource-rational contractualism: A triple theory of moral cognition. *Behavioral and Brain Sciences*, 1–38.

Michael, J., Sebanz, N., & Knoblich, G. (2016). The sense of commitment: A minimal approach. *Frontiers in psychology*, *6*, 162497.

Michael, J., & Tummolini, L. (2025). Intrinsically motivated norm compliance and the sense of obligation. *Current Opinion in Psychology*, *65*, 102043.

Theriault, J. E., Young, L., & Barrett, L. F. (2021). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, *36*, 100–136.