Engineering and reverse-engineering morality

Sydney Levine (smlevine@mit.edu)[1,2], Fiery Cushman[2], Iyad Rahwan[3], Joshua Tenenbaum[1]
[1]Department of Brain and Cognitive Sciences, MIT, 43 Vassar Street, Cambridge, MA USA
[2]Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138 USA
[3]Center for Humans & Machines, Max Planck Institute for Human Development, Lentzealle 94, Berlin 14195, Germany

Keywords: moral judgment and decision-making; AI ethics

Overview

Recent years have witnessed a burst of progress on building formal models of moral decision-making. In psychology, neuroscience and philosophy, the goal has been to "reverse-engineer" the principles of human morality. Meanwhile, in AI ethics, the goal has been to engineer systems that can make moral decisions, in some ways inspired by how humans do this. We aim to showcase the state of the art in both fields and to show how they can be hybridized into a computational cognitive science of morality.

Workshop Structure

This workshop would span a full day, split into two sessions. Each session will be composed of four talks and a discussant. The first session, "Reverse-Engineering the Morality of Humans", will focus on the ways that human moral judgment can be studied using a reverse-engineering approach. Talks in this segment will present research using computational tools, formal modeling, game theoretical approaches, and/or a rational analysis framework. The second session, "Learning from humans to build moral AI", will showcase a series of proposals for building ethical AI that draw insights from cognitive science. Talks in this segment look to how human cognition navigates the complex moral world as a starting place to generate engineering solutions to similar problems. A poster session will be held between the two segments, highlighting recent work from emerging scholars on either theme. Posters will be presented (along with informal socializing) on gather.town. We will end the workshop with a panel discussion between all speakers and organizers.

We have invited eight speakers (whose contributions are described in detail below) and will widely advertise a call for submissions, planning to select two additional speakers and a series of posters.

Diversity Statement: Our speakers come from a wide range of disciplines within cognitive science (computer science, AI development, philosophy, evolutionary psychology, social psychology, and law), use a range of methods (game theoretic analysis, behavioral techniques, engineering approaches, and developmental methods), represent a range of career stages and ages (from first-year graduate students to full professors) and is balanced on gender. However, our speakers do not represent the full range of racial diversity that is present in cognitive science (all but one of our speakers identifies as White) and nearly all of our speakers are cisgender. Therefore, in selecting additional contributed talks, we will prioritize increasing representation from under-represented groups in order to broaden the range of perspectives and voices contributing to this workshop.

Reverse-Engineering the Morality of Humans

This session showcases recent progress in "reverse engineering" the computational basis of human morality.

Jean-Baptiste André and Nicolas Baumard use evolutionary game-theoretic approaches to provide an analysis of the problem that human moral systems are designed to solve: how to ensure mutual benefit between cooperators (Baumard, André, & Sperber, 2013). They argue that the moral sense calculates the opportunity costs paid by others when they cooperated, so as to be able to reward them appropriately later. Building on this, **Shaun Nichols**'s work describes the form of the representational structures used to do this: moral rules. Nichols uses a Bayesian learning framework to investigate how we infer the scope of norms both experimentally and computationally and finds that people infer the contents of norms (and to whom they apply) in rationally appropriate ways (Nichols, 2021). Bridging the ideas of the first two speakers, **Sydney Levine, Fiery Cushman, and Joshua Tenenbaum** address the fundamental puzzle of how moral rules seem to be both rigid and flexible. The authors draw on the philosophical tradition of contractualism to describe how novel rules can be inferred by considering what everyone in the situation would contract to (Levine, Kleiman-Weiner, Schulz, Tenenbaum, & Cushman, 2020).

As commentator, **Gillian Hadfield** argues that the appropriate unit of analysis to understand morality is not a particular norm or behavior (as many of the speakers suggest); instead, it is the system that can support the establishing of particular norms or behaviors. Rather than asking, how do humans think or choose in moral ways we should ask, how do humans sustain systems of morality (Hadfield & Weingast, 2014)? Linking this session to the next, Hadfield then presents her own work using computational methods and a multi-agent reinforcement learning systems, which suggests a way to build generalizable

machine learning models that are capable of participating in a normative system with humans.

Learning from humans to build moral AI

As AI systems are developed to deal with complex moral scenarios, they are beginning to look towards cognitive science to see how the human mind solves similar challenges. This session explores the question of how AI systems can learn from and adapt to human norms and preferences under uncertainty.

Dylan Hadfield-Menell and Stuart Russell open the session by proposing that AI must abandon its "standard model" in which machines optimize a fixed, known objective – moral or otherwise. Instead, machines should be designed to act in ways that further human interests, while remaining fundamentally uncertain about what those interests are. The approach is formally instantiated as "assistance games" and shown to have a number of desirable properties not shared by the standard model. The authors look to moral philosophy for insights concerning how preferences aggregate across individuals, the plasticity of preferences, and how to handle relative preferences (Hadfield-Menell, Dragan, Abbeel, & Russell, 2016; Russell, 2020). Building on this, **Alison Gopnik** proposes that we can look to cognitive science for inspiration on how to infer and aid the interests of humans; this after all, is the role of a human caretaker. A parent is a person whose self has been expanded to include the values and interests of another agent, even when those values and interests are different from his, and even when that agent is capable of inventing new goals and values to suit new circumstances (Gopnik, 2016). Describing the computational mechanisms of care can help us develop morally competent AI. **Peter Railton** looks to the moral development of human infants for a model of how largely unsupervised learning processes for understanding and participating in social interactions could contribute both to the development of general-purpose intelligence and a capacity for distinctively moral evaluation and action. Core elements of AI ethics, then, might not need to be "built in" so much as acquired, an acquisition process that, as AI agents interact with humans and one another, could lead to the development of social-contract reasoning capacities (Railton, 2020).

As commentator, **Sholei Croom** argues for the virtue of maintaining higher levels of nuance in our models of morality than cognitive science tends to allow. Croom points out that while computational models have had success producing accurate predictions of individual human behavior in isolated experimental settings, this modeling strategy tends to fail when applied in broader social contexts. Computational models risk entrenching an implicit assumption that cognition and behavior manifest in a single universal way, and in doing so neglect the many complex factors that contribute to social behavior such as structural power and historical contingency.

Finally, there will be a closing panel discussion of all organizers and presenters, with **Iyad Rahwan** joining to talk about societal implications of engineering and reverse engineering morality.

Presenters and Organizers

Jean-Baptiste André is a research fellow at the French National Centre for Scientific Research (CNRS), Department of Cognitive Sciences. He will present his work with **Nicolas Baumard**, Research Director at the CNRS and Professor at the Ecole Normale Supérieure.

Sholei Croom is a graduate student in Psychological and Brain Sciences at Johns Hopkins.

Alison Gopnik is a Professor of Psychology and an affiliate Professor of Philosophy at UC Berkeley.

Gillian Hadfield is a Professor of Law and Strategic Management at U. of Toronto.

Dylan Hadfield-Menell, a graduate student in Electrical Engineering and Computer Sciences at UC Berkeley, will present his work with **Stuart Russell**, a Professor of Computer Science and Engineering at UC Berkeley.

Sydney Levine, a postdoc in Psychology at Harvard and in Brain and Cognitive Sciences at MIT, will present her work with **Fiery Cushman**, an Associate Professor in Psychology at Harvard and **Joshua Tenenbaum**, a Professor in Brain and Cognitive Sciences at MIT.

Shaun Nichols is a Professor of Philosophy at Cornell.

Iyad Rahwan is a Director of the Max Planck Institute for Human Development.

Peter Railton is a Professor of Philosophy at U. of Michigan.

References

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behav. and Brain Sci.*, *36*(1), 59.

Gopnik, A. (2016). *The gardener and the carpenter.* Macmillan.

Hadfield, G. K., & Weingast, B. R. (2014). Microfoundations of the rule of law. *Annual Review of Political Science*, *17*, 21–42.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. In *Proceedings of the 30th international conference on neural information processing systems* (pp. 3916–3924).

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158–26169.

Nichols, S. (2021). *Rational rules: Towards a theory of moral learning*. Oxford University Press, USA.

Railton, P. (2020). Ethical learning, natural and artificial. In *Ethics of artificial intelligence* (pp. 45–78). Oxford University Press.

Russell, S. (2020). *Human compatible: Ai and the problem of control*. Penguin Books.