Computational Frameworks for Modeling Moral Cognition and Cooperation

Arthur Le Pargneux (arthurlepargneux@fas.harvard.edu)[1]
[1]Department of Psychology, Harvard University

Max Kleiman-Weiner (maxkw@uw.edu)[2]
[2]Foster School of Business & Paul G. Allen School of Computer Science and Engineering, University of Washington

Maximilian Maier (max.maier@wbs.ac.uk)[3]
[3]Behavioural Science Group, Warwick Business School, University of Warwick

Julien Lie-Panis (jliep@protonmail.com)[4]
[4]Max Planck Institute for Evolutionary Biology

Joshua B. Tenenbaum (josh.tenenbaum@gmail.com)[5]
[5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Keywords: moral cognition; cooperation; cognitive modeling; game theory; reinforcement learning; Bayes; Theory of Mind

Overview

There is widespread agreement across disciplines that morality has evolved to facilitate cooperation among agents with conflicting interests. Traditional approaches to the study of morality and cooperation either abstain from formal models (e.g., moral psychology in social psychology) or model cooperation without much concern for underlying psychological processes (e.g., evolutionary models of reciprocity, kin selection etc.). Computational cognitive scientists are starting to enrich classic computational frameworks (Bayesian inference, reinforcement learning, game theory, expected utility) with psychological mechanisms, traits, and constraints (Theory of Mind, metacognition, trust, joint planning, resource rationality) to model behavior, learning and evolutionary processes, and develop a finer understanding of the cognitive underpinnings of morality and cooperation. This progress is accelerated by current advances in modeling (probabilistic programming languages specialized for social cognition, large-language models, multi-agent RL) and computational efficiency (GPUs, numerical computation). Building formal models of moral cognition and cooperation that are both computationally precise and psychologically rich opens the door for cumulative progress in our understanding of their principles, origins, and functioning, and for the engineering of cooperative and ethical capabilities in artificial systems.

Contributors

This symposium brings together researchers from multiple disciplines (experimental psychology, evolutionary biology, artificial intelligence, economics, cognitive anthropology) who combine various computational approaches to build psychologically rich models of morality and cooperation.

Max Kleiman-Weiner (Washington), **Alejandro Vientós** (MIT), **David Rand** (Cornell), and **Josh Tenenbaum** (MIT) bring Theory of Mind and Bayesian inference processes into evolutionary models to understand the emergence of direct and indirect reciprocity (Kleiman-Weiner et al., 2025).

Maximilian Maier (Warwick), **Vanessa Cheung** (UCL), and **Falk Lieder** (UCLA) discuss moral decisions in terms of strategy selection (Maier et al., in press) and use RL models to investigate how metacognition and learning processes shape the trade-off between rules and cost-benefit reasoning in moral decision-making (Maier et al., 2025).

Julien Lie-Panis (Max Planck Institute for Evolutionary Biology), **Léo Fitouchi** (Institute for Advanced Study in Toulouse, Toulouse School of Economics), **Nicolas Baumard** (CNRS, ENS Paris), and **Jean-Baptiste André** (CNRS, ENS Paris) incorporate trust into evolutionary game theoretic models to clarify the adaptive logic of rigid moral rules, explaining when and why they are favored over flexible cost-benefit reasoning (Lie-Panis et al., 2025).

Arthur Le Pargneux (Harvard), **Sydney Levine** (NYU), **Hossam Zeitoun** (Warwick), **Nick Chater** (Warwick), **Josh Tenenbaum** (MIT), and **Fiery Cushman** (Harvard) study how people sustain coordination in changing environments by developing a model centred around joint planning which combines principles of utility maximization, Bayesian inference, and resource rationality (Le Pargneux et al., 2026).

The panel discussion will be led by **Josh Tenenbaum**.

Evolving general cooperation with a Bayesian theory of mind

Kleiman-Weiner, Vientós, Rand, Tenenbaum

Theories of the evolution of cooperation through reciprocity explain how unrelated self-interested individuals can accomplish more together than they can on their own. The most prominent theories of reciprocity, such as tit-for-tat or win-stay-lose-shift, are inflexible automata that lack a theory of mind—the human ability to infer the hidden mental states in others' minds. Here, we develop a model of reciprocity with a theory of mind, the Bayesian Reciprocator. We evaluate the Bayesian Reciprocator using a generator over games where every interaction is unique, as well as in classic

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

environments such as the iterated prisoner's dilemma. The Bayesian Reciprocator enables the emergence of both direct-reciprocity when games are repeated and indirect-reciprocity when interactions are one-shot but observable to others. In an evolutionary competition, the Bayesian Reciprocator outcompetes existing automata strategies and sustains cooperation across a larger range of environments and noise settings than prior approaches. This work quantifies the advantage of a theory of mind for cooperation in an evolutionary game theoretic framework and suggests avenues for building artificial agents with more human-like learning mechanisms that can cooperate across many environments.

Learning from outcomes shapes reliance on moral rules versus cost-benefit reasoning

Maier, Cheung, Lieder

Real-world moral decisions are constrained by limited information and bounded cognitive resources, necessitating heuristic strategies. We argue that choices in moral dilemmas should be analyzed in terms of decision strategies rather than ethical theories, and outline general principles that shape strategy selection (bias variance trade-off, opportunity cost of thinking). Further, we show empirically how people learn which decision strategies to rely on in practice. Specifically, we focus on disagreements between moral rules and 'utilitarian' cost–benefit reasoning (CBR). In a new paradigm, participants (N = 2,328) faced dilemmas between one choice prescribed by a moral rule and one by CBR. They observed the consequences of their decision before the next dilemma. Across 4 experiments, we found adaptive changes in decision-making over 13 choices: participants adjusted their decisions according to which decision strategy (rules or CBR) produced better consequences. Using computational modelling, we showed that many participants learned about decision strategies in general (metacognitive learning) rather than specific actions. Their learning transferred to incentive-compatible donation decisions and moral convictions beyond the experiment. We conclude that metacognitive learning from consequences shapes moral decision-making and that individual differences in morality may be surprisingly malleable to learning from experience.

Why human societies adopt rigid moral rules: The efficiency-robustness tradeoff

Lie-Panis, Fitouchi, Baumard, André

Humans are capable of remarkably flexible moral judgment. Yet societies rely on rigid rules—obligations and prohibitions that apply categorically, even when case-by-case reasoning could yield better outcomes. Why would a species capable of such flexibility bind itself to inflexible rules? We propose that rigid rules arise as social technologies for managing ambiguity around noncooperation. People often have legitimate reasons for failing to cooperate, yet those reasons are typically opaque to observers, allowing opportunists to disguise selfishness as justified hardship. We formalize this idea with an evolutionary game-theoretic model. Two cooperative equilibria emerge: a flexible norm that accommodates legitimate excuses but is vulnerable to exploitation, and a rigid norm that closes this loophole by mandating cooperation even when inefficient. Findings reveal an efficiency–robustness trade-off: flexibility maximizes welfare when trust is secure, whereas rigidity preserves cooperation when trust is fragile. This explains why rigid rules prevail in low-trust settings—interactions with strangers, formal institutions, or tight societies—where flexibility is more common in high-trust contexts.

Balancing precedent and mutual benefit in human coordination

Le Pargneux, Levine, Zeitoun, Chater, Tenenbaum, Cushman

Human coordination depends on two complementary mechanisms: forward-looking strategies that enable flexible adaptation to new circumstances, and backward-looking mechanisms that rely on precedent and rule-following. Most models of coordination emphasize one mechanism or the other—either explaining how equilibria emerge and persist based on past experience, or how agents creatively generate novel solutions to achieve anticipated mutual benefit—but not how the two interact. Here we introduce a cognitive model and experimental paradigm to capture the dynamics of both processes and, crucially, the arbitration between them. In two preregistered experiments (n = 510; 30,420 choices), participants repeatedly solve coordination problems that can be addressed either by generalizing past solutions or by adopting novel ones when precedent becomes inefficient. This design allows us to examine the conditions under which individuals or dyads decide to abandon entrenched equilibria and transition to novel coordination solutions by arbitrating between mutual benefit and precedent. By formally modeling both forward- and backward-looking mechanisms, and their arbitration, we provide a unified framework for understanding how coordination can be both stable and adaptable—a property that underlies everyday cooperative behavior, social norms, and institutional evolution.

References

Kleiman-Weiner, M., Vientós, A., Rand, D. G., & Tenenbaum, J. B. (2025). Evolving general cooperation with a Bayesian theory of mind. *Proceedings of the National Academy of Sciences*, *122*(25), e2400993122.

Le Pargneux, A., Levine, S., Zeitoun, H., Chater, N., Tenenbaum, J. B., & Cushman, F. (2026). Balancing precedent and mutual benefit in human coordination. *OSF*.

Lie-Panis, J., Fitouchi, L., Baumard, N., & André, J.-B. (2025). Why human societies adopt rigid moral rules: The efficiency–robustness trade-off. *OSF*.

Maier, M., Cheung, V., & Lieder, F. (2025). Learning from outcomes shapes reliance on moral rules versus cost–benefit reasoning. *Nature Human Behaviour*.

Maier, M., Cheung, V., & Lieder F. (in press). Moral decision-making with bounded cognitive resources and limited information. *Trends in Cognitive Sciences.*