disapproval, coordination failures, or even legal consequences if the norm is of a more formal kind.

In this commentary, I propose that everyday moral cognition classifies certain situations as *rule-governed*. In such contexts, people expect behaviour to be guided by a standing – or hypothesised – norm rather than by on-the-fly preference trading or interest guessing. Once an interaction is classified this way, agents either look outward – invest effort in discovering the actual rule – or look inward – simulate a plausible rule, for instance by using the generalisation machinery the Target Article describes.

A central agenda for future research, therefore, is to chart the situational cues that trigger the "rule-governed" classification. Two candidate cues stand out and lend themselves to laboratory and field tests:

Coordination incentives. When the cost of miscoordination is high – as in road traffic, air-traffic control, or medical triage – people intuitively expect a focal point, naturally supplied by a rule. In such domains, relying on ad hoc preference trading is reckless because every actor's payoff hinges on seamless alignment. Experiments can vary coordination stakes (e.g., low- vs high-penalty traffic simulations) and observe whether participants shift from negotiating outcomes to searching for, or inventing, traffic – code – like rules.

Private-public gradient. In the *private* sphere – for example, interactions with family and friends – we usually know others' idiosyncratic preferences and expect our conduct to be judged by the care and concern we show, not by blind rule adherence. By contrast, the *public* sphere is marked by anonymity, formality, and a broader audience; here personalised knowledge is scarce (second-guessing strangers' preferences may even be frowned upon) and evaluations hinge on visible rule compliance ("no talking in libraries," "queue here"). The more public, formal, or unfamiliar the setting, the more people default to rule-based reasoning. Researchers can operationalise "publicness" through factors such as audience size, social distance, or institutional décor, and then measure whether these manipulations push decisions from preference-sensitive bargaining to explicit or hypothesised rule following.

Identifying and testing such cues will reveal how often – and how quickly – agents flip from preference-tracking to rule-tracking procedures, thereby refining the virtual-bargaining framework.

A second frontier for the framework concerns *what* agents think they are doing once they recognise a context as rule-governed. The Target Article does not distinguish between actors who *believe a rule already exists* (and therefore try to discover it) and actors who merely *conjecture what a rule could be*. Intuitively, people invest more effort in learning and compliance when they treat the rule as real, and excuse themselves more readily when the rule is only hypothetical. Parsing this difference could clarify why some emergent norms harden into law whereas others remain provisional conventions – a distinction the resource-rational model is well placed to explain but has yet to explore.

The resource-rational contractualist model proposed by Levine and colleagues deftly explains how agents settle on the right course of action in interactive settings with limited information and communication. As I argued in this commentary, its explanatory reach will grow once it tackles two further questions: when do agents classify an encounter as rule-governed, and how do they treat the relevant rule – as something to be discovered or merely hypothesised? Mapping the situational cues that steer these judgements – especially high coordination incentives and the private-public gradient – and tracing their downstream effects on learning, compliance, and norm crystallisation appears to be a promising avenue for future research. In short, more empirical work is needed to establish which contexts people regard as rule-governed, and the Target Article offers a robust theoretical scaffold for that agenda.

Acknowledgements. I am grateful to Josh Knobe, Shaun Nichols, Revati Shivnekar, and Kacper Poradzisz for conversations that helped clarify the ideas developed in this commentary.

Financial support. This research was supported by the National Science Centre, Poland (NCN), grant 2022/47/B/HS5/03415.

Competing interests. I declare no competing interests.

Contractualist moral cognition needs a legalese-of-thought

Kartik Chandra[a]* and Max Kleiman-Weiner[b]

[a]Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, USA and [b]University of Washington, 1410 NE Campus Parkway, Seattle, WA, USA
kach@mit.edu
maxkw@uw.edu

*Corresponding author.

doi:10.1017/S0140525X25101726, e137

Abstract

What exactly *is* a contract? In this commentary, we discuss the need for a formal theory of how contracts are represented in the mind. Drawing inspiration from the representation of real-world contracts, we suggest that *probabilistic programs* are an appealing candidate representation.

The notion of a *contract* plays a central role in the resource-rational contractualism framework. But what exactly *is* a "contract"? How are social contracts represented in the mind?

In this commentary, we discuss the need for a formal theory of contract representations. Such a theory would allow us to extend Levine et al.'s account of moral cognition in many valuable directions. First, an explicit account of contract representations would lay the groundwork for precisely specifying a computational architecture for how moral agents encode, store, and manipulate contracts. What computational operations support evaluating, comparing, composing, writing, and updating contracts? How do our representations of contracts interface with cognitive systems for planning, inference, and decision-making? How do we learn to work with contracts, and how do we incorporate past moral experiences ("precedent" and "case law") into our future contractual reasoning (Feng et al., 2023)?

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Second, a formal treatment of contract representations would help us understand how moral agents manage resource-rational trade-offs. Different representational schemes come with different computational costs and benefits. Positing explicit data structures and algorithms for working with contracts would allow us to make quantitative, testable predictions about moral cognition: for example, the threshold where people might switch from converging on a contract specified as a series of special cases to a contract specified as a general rule.

Third, a formal account of contract representations would allow us to study how people come to acquire contracts: either by writing new contracts in new situations or by inferring existing contracts from observation. Contract acquisition can be modeled as searching over a space of possible contracts. For example, Oldenburg and Zhi-Xuan (2024) propose a model of how observers might infer a society's norms by observing agents' (norm-constrained) actions and then performing Bayesian inference over a space of possible norms. A formal account of contract representations would specify more generally what the search space is for contract creation and inference.

What might this general search space look like? In Oldenburg and Zhi-Xuan's model, the space of possible norms is generated by a grammar of first-order predicates over permissible behavior. More generally, a combinatorial language-like representation for contracts is appealing, as it provides a compositional approach to building complex representations from simple components. The mind might recombine "clauses" from prior agreements to create new contracts in new situations. One possibility, then, is that contracts are represented as *programs* in a language-of-thought for agreements: what we might call the "legalese-of-thought."

It is indeed quite natural to conceive of contracts as programs, and recent years have seen many efforts to use specialized programming languages to represent legal agreements as executable "smart contracts." Because smart contracts are formalized unambiguously in code, they enjoy the benefits of transparency, determinism, and decentralization – at least in principle. In practice, however, this formality comes at a cost: smart contracts famously struggle to handle the ambiguity inherent in the real world. They cannot adapt to new, unanticipated situations, and when edge cases inevitably occur, they cause millions of dollars in losses (Atzei, Bartoletti, & Cimoli, 2017). This brittleness is inconsistent with the remarkable flexibility of the human moral mind, which suggests that deterministic programs are not the answer after all.

What are we missing here? An important observation is that legal contracts rarely, if ever, explicitly specify responses to every possible contingency, but the remaining ambiguity is often considered a "feature, not a bug" (Hadfield-Menell & Hadfield, 2019). For example, treaties between nations sometimes contain vague terms to facilitate agreement while deferring contentious specifics to future negotiation (Koremenos, 2016). Similarly, constitutional principles are frequently articulated at a high level of abstraction to accommodate evolving societal values and unforeseen technological development (Sunstein, 1996). The Fourth Amendment of the US Constitution prohibits "unreasonable searches and seizures" but leaves "unreasonable" undefined (see, e.g., Katz v. United States), and the Clean Air Act requires protecting public health with an "adequate margin of safety" to address uncertainties associated with hazards that have not yet been fully identified (McClellan, 2012). The wisdom in such agreements is in acknowledging the many sources of uncertainty inherent in moral reasoning: from uncertainty about the current state of the world to uncertainty about future outcomes to uncertainty about the meaning of the contract itself.

If principled reasoning under uncertainty really is a fundamental feature of contract representations, then *probabilistic* programs may offer a more compelling representational framework than deterministic programs. Probabilistic programs combine the strengths of explicit symbolic reasoning with the flexibility of graded statistical inference. In particular, probabilistic programming languages like memo (Chandra et al, 2025), which center theory-of-mind reasoning, can be used to study how moral agents come up with contracts in the face of uncertainty over each other's beliefs, desires, and intentions. For example, when deciding whether to take an action that affects another person without their consent (Levine et al, 2024), a moral agent may consider a contract that requires making graded probabilistic inferences about the other party's values ("it is okay if I am confident that this is what you want"). To study how moral agents interpret written rules, we might follow Wong et al. (2023) and model natural-language rules as being translated into a probabilistic legalese-of-thought by a language model. As for *unwritten* rules, we might consider how uncertainty creates space for moral agents to coordinate contracts around Schelling points: salient equilibria that emerge without explicit communication (Schelling, 1960).

Effective moral cognition demands representations that are explicit enough to coordinate social inference and behavior, yet flexible enough to accommodate the pervasive uncertainty we have about our world, our future, and each other. A probabilistic "legalese-of-thought" would reveal the internal data structures that make our moral inferences tractable, quantify the resource costs that shape moral reasoning, and provide a common language for cognitive scientists, computer scientists, and philosophers to specify and test theories of moral agreement.

Acknowledgements. None.

Financial support. KC was supported by the Hertz Foundation and the NSF Graduate Research Fellowship Program. MKW was supported by the Cooperative AI Foundation, the Foresight Institute, and the Sony Research Award Program.

Competing interests. None.

References

Atzei, N., Bartoletti, M., & Cimoli, T. (2017). A survey of attacks on Ethereum smart contracts. In *International conference on principles of security and trust.* (pp. 164–186). Berlin, Heidelberg: Springer Berlin Heidelberg.

Chandra, K., Chen, T., Tenenbaum, J. B., & Ragan-Kelley, J. (2025). *A Domain-Specific Probabilistic Programming Language for Reasoning about Reasoning (or: a memo on memo).* https://doi.org/10.31234/osf.io/pt863

Feng, K. J., Chen, Q. Z., Cheong, I., Xia, K., & Zhang, A. X. (2023). Case repositories: Towards case-based reasoning for AI alignment. *arXiv preprint arXiv:2311.10934.*

Hadfield-Menell, D., & Hadfield, G. K. (2019). Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 417–422).

Katz v. United States, 389 U.S. 347 (1967).

Koremenos, B. (2016). *The continent of international law: Explaining agreement design*. Cambridge University Press.

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J. B. (2024). When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *Cognition*, *250*, 105790.

McClellan, R. O. (2012). Role of science and judgment in setting national ambient air quality standards: how low is low enough?. *Air Quality, Atmosphere & Health*, *5*(2), 243–258.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Oldenburg, N., & Zhi-Xuan, T. (2024). Learning and sustaining shared normative systems via Bayesian rule induction in Markov games. *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 23, 1510–1520.
Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press.
Sunstein, C. R. (1996). Social norms and social roles. *Columbia Law Review*, *96*, 903.
Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. arXiv preprint arXiv:2306.12672.

Moral decision-making entails negotiation over the psychological mechanisms underlying decisions

Paul Conway*, Jason Lam and Constanine Sedikides

B44 University Rd, University of Southampton, Southampton, UK
p.conway@soton.ac.uk
c.sedikides@soton.ac.uk
Jason.Lam@soton.ac.uk

*Corresponding author.

doi:10.1017/S0140525X25101659, e138

Abstract

Levine et al. offer a compelling theory of moral decision-making, yet underemphasize the role of social perception. We argue that moral judgments function not only to evaluate actions but also to signal moral character. This signaling shapes social interactions and guides moral processing, highlighting the need to integrate reputational and relational dynamics into models of moral cognition.

We commend Levine and colleagues on proposing an ambitious and integrative theory of moral decision-making. We find ourselves in broad agreement with the theory, although we argue that it does not go far enough. The theory focuses on (implied or actual) negotiation over elements including which moral standards to employ and whether to judge outcomes versus actions. It neglects, however, a crucial aspect of human moral psychology: moral judgments – and the cognitive processes underlying them – function not only as an evaluation of actions or events but also as signals of an individual's moral character and social reputation. This signaling function introduces important strategic considerations, as individuals must navigate what their morally relevant thoughts, feelings, and behaviors convey to themselves and others in complex social environments.

Levine and colleagues briefly acknowledge the role of social evaluation in moral judgment, stating that "People do not just evaluate acts, consequences, and agreements, but also other people" (sect. 6, para 2). However, this phrasing arguably understates the centrality of social perception and meta-perception in moral decision-making. Individuals do not merely infer traits – such as warmth, competence, and moral character – from others' moral decisions (Rom, Weiss, & Conway, 2017); these inferences meaningfully shape downstream social behaviors, including trust and social role selection (Everett et al., 2016). Yet, individuals make more than trait-level attributions: they also infer the underlying mental states or "mental occurrents" driving moral decisions, such as the degree of concern for victims or commitment to moral rules (Critcher, Helzer, & Tannenbaum, 2020). Such inferences carry distinct social consequences. For example, individuals who reject sacrificial harm based on rigid rule adherence are often perceived as more self-righteous than those who reject it due to emotional concern for victims, despite endorsing the same moral judgment (Weiss et al., 2024).

Moreover, individuals demonstrate metacognitive sensitivity: they understand how different moral judgments appear to others, strategically adjusting their moral judgments to imply particular traits or decision styles in line with perceived expectations in a given context (Rom & Conway, 2018). Hence, moral judgments do not merely reflect deliberation over optimal strategies for determining right action or desirable outcomes; they also function as negotiations of moral selves, shaped through the communication and inference of underlying psychological constructs or mental occurrents. Individuals care deeply about their moral self-image and reputation (Heiphetz et al., 2018), and they strategically use moral decision-making and moral communication to bolster this self-concept (Rothschild & Keefer, 2017; Brady & Van Bavel, 2025). Given the ubiquity of self-comparison (Gerber, Wheeler, & Suls, et al., 2018), the presence of ostensibly more virtuous others can pose a threat to one's moral self, eliciting shame or defensiveness (Monin, Sawyer, & Marquez, 2008). Accordingly, the processes described by Levine et al. triple theory of moral cognition must also account for the reputational and relational dimensions of moral judgment. That is, moral evaluations are influenced not only by (implied or actual) negotiation over the appropriateness of moral considerations themselves but also by the perceived implications of how and why one considers such things for one's moral self and the moral selves of others.

Considering how individuals interpret the psychological mechanisms behind others' moral decisions may help to clarify when and why they prefer one mode of reasoning over another. Levine and colleagues hypothesize that "the higher the stakes of a moral decision, the greater the cost of a suboptimal choice. Thus, like unique situations, high-stakes situations should favor those mechanisms of moral judgment that achieve greater accuracy at the cost of effort" (sect 5.1.1, para 5). Although this hypothesis holds in some domains, it does not fully account for findings pertaining to high-stakes moral dilemmas – such as life-and-death scenarios – where responses often reflect heuristic adherence to deontological rules rather than calculated utilitarian reasoning (Bostyn, Roets, & Conway, 2022; Piazza & Landy, 2013). However, these seemingly suboptimal responses may serve an important social function: they signal traits such as warmth, reliability, and trustworthiness, which are attributes of a desirable, cooperative partner (Everett, Pizarro, & Crockett, 2016; Sacco & Brown, 2017).

From a functionalist perspective – one broadly aligned with the authors' approach – the attribution of underlying mechanisms in others' moral judgments carries socially relevant information about