Resource-rational moral judgment

Sarah A. Wu[1], Xiang Ren[2, 3], Tobias Gerstenberg[1], Yejin Choi[2, 4], Sydney Levine[2]
sarahawu@stanford.edu
[1]Department of Psychology, Stanford University
[2]Allen Institute for Artificial Intelligence
[3]Department of Computer Science, University of Southern California
[3]Department of Computer Science, University of Washington

Abstract

There is wide agreement that the mind has different mechanisms it can use to make moral judgments. But how does it decide which one to use when? Recent theoretical work has suggested that people select mechanisms of moral judgment in a way that is *resource-rational* — that is, by rationally trading off effort against utility. For instance, people may follow general rules in low-stakes situations, but engage more computationally intensive mechanisms such as consequentialist or contractualist reasoning when the stakes are high. Here, we evaluate whether humans and large language models (LLMs) exhibit resource-rational moral reasoning in two moral dilemmas by manipulating the stakes of each scenario. As predicted, we found that the higher the stakes, the more people employed a more effortful mechanism over following a general rule. However, there was mixed evidence for similar resource-rational moral reasoning in the LLMs. Our results provide evidence that people's moral judgments reflect resource-rational cognitive constraints, and they highlight the opportunities for developing AI systems better aligned with human moral values.

Keywords: moral judgment; resource rationality; large language models

Introduction

Moral philosophers and psychologists have long wrestled with how to reconcile conflicting normative theories and scattered descriptive accounts of moral cognition. One set of accounts, *consequentialism*, involves reasoning about and maximizing consequences such as welfare (e.g., Mill, 1879). Another set of accounts, *deontology*, concerns the permissibility of actions according to certain properties they have or to moral rules, regardless of the consequences (e.g., Kant, 1785). Finally, *contractualist* accounts argue that moral permissibility depends simply on what affected parties would agree to (e.g., Rawls, 1971; Scanlon, 2000). These accounts are not only normative theories, but have also been separately shown to guide people's actual moral judgments (for a review, see Levine et al., 2023).

Given evidence for all of these moral mechanisms at play, how do people decide which mechanism to deploy and when? Recent theoretical work has proposed that people select the appropriate moral mechanism to use in a way that is *resource-rational* — that is, by efficiently trading off cognitive effort against precision or payoff (Levine et al., 2023). The idea that people make rational use of limited cognitive resources has been successfully applied to many cognitive processes, including memory (e.g., Anderson & Milson, 1989), reasoning (e.g., Icard & Goodman, 2015; Lieder et al., 2018), and decision-making (e.g., Vul et al., 2014). We adopt the term *resource rationality* (Lieder & Griffiths, 2020), although other names have been used such as *bounded optimality* (Horvitz, 1987; Russell & Subramanian, 1995) and *computational rationality* (Gershman et al., 2015; Lewis et al., 2014).

In this paper, we investigate whether resource rationality may govern the mechanisms of moral judgment used by people and state-of-the-art large language models (LLMs). As LLMs and other AI systems become increasingly integrated into society and begin to make morally-charged decisions, it is important to gauge their alignment with the diversity and flexibility of human moral reasoning. LLMs can often produce the "right" answer and score high on moral benchmarks, especially with fine-tuning (e.g. Jiang et al., 2022) and careful prompting (e.g. Jin et al., 2022). However, it is critical to understand how closely they mimic the underlying *mechanisms* of human moral cognition in order for us to understand and predict how they might behave in completely novel contexts. Resource rationality provides a useful window to gain insight into underlying processes. Therefore, alongside our investigation of whether and how human morality adheres to resource rationality, we also test whether the judgments of LLMs follow this framework as well.

Moral reasoning in humans

Prior work shows that people use a variety of cognitive mechanisms to make moral judgments, from consequentialist, to deontology, to contractualist processes. Many studies show that people consider and tradeoff consequences like welfare in moral dilemmas such as the trolley problem (e.g., Cushman et al., 2006; Greene, 2013). But people are also sensitive to intrinsic properties of actions, such as whether they were intended or involve personal force, (Cushman et al., 2006; Greene et al., 2009), or violate moral rules (Haidt, 2001; Nichols & Mallon, 2006), regardless of the outcomes. In some cases, people's judgments are best explained not by welfare-based or action-based mechanisms, but by other mechanisms such as universalization (Kwon et al., 2023; Levine et al., 2020) and virtual bargaining (André et al., 2023; Baumard et al., 2013; Everett et al., 2016; Levine et al., 2022) that share more undertones with contractualist theories.

What this wealth of research highlights is that the human moral mind is incredibly flexible. Much of the time we follow rules, which are powerful and general, but we also often

Table 1: **Experiment items.** Questions given to participants and LLMs in each scenario. All participants first read the vignette and then answered the three questions on separate pages. In the *triage* scenario, rule-breaking reflects using a consequentialist mechanism, and in the *property* scenario, rule-breaking reflects using a contractualist (virtual bargaining) mechanism.

Question	*Triage* scenario	*Property* scenario	Response format
rule-following	How morally acceptable is it for the doctor to treat the first person in line first?	How morally acceptable is it for Hank to turn the stranger away?	Likert scale from "very unacceptable" (1) to "very acceptable" (7), with midpoint "neutral/unsure"
rule-breaking	How morally acceptable is it for the doctor to treat someone other than the first person in line first?	How morally acceptable is it for Hank to carry out the stranger's request?	Likert scale from "very unacceptable" (1) to "very acceptable" (7), with midpoint "neutral/unsure"
free-response	Who should the doctor ideally treat first?	Suppose that Hank accepts the stranger's offer. How much money, if any, should he give to each neighbor?	open-ended text box

exempt and continuously revise them. Prior work investigating judgments about the moral acceptability of rule-breaking, such as cutting in line or violating another's property, suggests that many of the mechanisms described above are recruited when reasoning about rule-breaking (Awad, Levine, Loreggia, et al., 2022; Kwon et al., 2023; Levine et al., 2022).

Moral reasoning in LLMs

The flexibility of cognitive mechanisms involved in human moral reasoning poses an engineering challenge for LLMs. Researchers have sought to develop more moral AI systems alongside better understanding human morality in an emerging movement called *computational ethics* (Awad, Levine, Anderson, et al., 2022). Many studies have evaluated the performance of LLMs on datasets of general moral questions, such as SCRUPLES (Lourie et al., 2021), ETHICS (Hendrycks et al., 2021), Social Chemistry (Forbes et al., 2020), Moral Stories (Emelin et al., 2021), and the Commonsense Norm Bank (Jiang et al., 2022). While extensive, these benchmarks lack structured experimental contrasts that can provide insight on underlying mechanisms. Others have used a more "machine psychology" approach (Binz & Schulz, 2023; Hagendorff, 2023) to probe the *why* and *how* of their responses, specifically investigating moral rule-breaking and action-based features in moral judgments (Almeida et al., 2024; Jin et al., 2022; Nie et al., 2023). Little work has explicitly studied other mechanisms of moral judgments, such as contractualist mechanisms, in LLMs.

Integrating mechanisms of moral judgment

What explains people's moral flexibility, and how do they integrate the plethora of moral mechanisms? Some prior work has suggested that deontological and consequentialist reasoning map onto "fast", intuitive model-free decision-making, and "slow", deliberate model-based decision-making, respectively (Crockett, 2013; Cushman, 2013; Cushman et al., 2010; Greene, 2013; Greene et al., 2004; but see De Neys, 2022). This *dual-process* theory posits that these two types of moral reasoning follow from different processes in the mind guided by different value representations, specifically action-based or outcome-based representations. Some research has found that, when under cognitive load such as time pressure, people make fewer consequentialist (Greene et al., 2008; Ham & van den Bos, 2010; Kroneisen & Steghaus, 2021) and more deontological judgments (Suter & Hertwig, 2011), supporting the dual-process theory. However, other studies have found no such effects (Tinghög et al., 2016) or even the opposite pattern (Hashimoto et al., 2022). Another theory that relates deontology and consequentialism is *threshold deontology*, which holds that deontology should be followed unless the consequences cross some threshold of badness, at which point consequentialism should be deployed instead (Moore, 1997). This theory finds empirical support (e.g., Ryazanov et al., 2023; Trémolière & Bonnefon, 2014), but is also not without its critiques (e.g., Alexander, 2000).

Both the dual-process theory and moderate deontology relate consequentialst and deontological mechanisms, but neither explains the role of contractualist mechanisms. Levine et al. (2023) propose a theory of resource-rational contractualism that integrates all three. They note that the function of morality is to guide people towards agreements of mutual benefit, and as such, contractualist mechansims may actually be the ideal approach to making moral judgments. They propose that other mechanisms, such as welfare-based and rule- or action-based ones, are simply efficient approximations of contractualism. Mechanisms that are more specific and accurate come at the cost of social and cognitive effort. Resource rationality posits that people select moral mechanisms by trading off the social and cognitive costs of engaging a mechanism against the mutual benefit it would achieve.

Overview of current paradigm

To investigate whether people and LLMs make such resource-rational trade-offs, we constructed two morally charged

scenarios—one involving medical triage and the other involving property violation—that could be handled using different mechanisms. In each scenario, we contrast two mechanisms in particular: one relatively simple (following a rule) and one more complex (calculating consequences or virtual bargaining). The scenarios were designed such that the use of each mechanism would produce a distinctive pattern of judgments and could therefore be differentiated. For instance, the *triage* scenario involves patients in line at an urgent care facility where there is a generally accepted rule for doctors to treat patients in the order that they arrive ("first-come, first serve", Pàmies et al., 2016). However, if doctors are trying to allocate their care most efficiently, then it will sometimes make sense to break the rule. Perhaps it would be acceptable to treat someone who arrived later but is in a critical life-threatening condition, or would be very quick to treat, over a patient who arrived earlier but has only minor injuries or needs a lengthy procedure. Considering these additional factors requires engaging a more effortful mechanism of moral reasoning that involves weighing and comparing outcomes.

Whether engaging a more complex mechanism is an efficient use of cognitive resources depends, among other factors, on the *stakes* of the situation. The higher the stakes, the greater the benefits of an optimal outcome and/or the greater the cost of a sub-optimal outcome, and thus the more we expect participants to use the more effortful-but-precise mechanism. Therefore, we can manipulate the stakes of each scenario and observe whether moral judgments shift according to resource-rational predictions. Stakes have been studied in the form of monetary incentives in tasks involving bargaining (e.g. Andersen et al., 2011; Cameron, 1999; Larney et al., 2019; Novakova & Flegr, 2013; Yamagishi et al., 2016), cheating (e.g. Kajackaite & Gneezy, 2017; Rahwan et al., 2018), and altruism (e.g. Burum et al., 2020). In the context of resource rationality, stakes have been manipulated to probe other cognitive processes such as reinforcement learning, where it was found that people used more resource-intensive learning strategies when the reward multiplier was higher (Kool et al., 2017).

Experiment

We designed two pairs of moral dilemmas manipulating the stakes of the situation in a 2×2 between-subjects design. The dilemmas could be solved by applying different moral mechanisms, some relatively simple and others more complex. One scenario involves medical triage and pits rule-following (simple) against consequentialist reasoning (complex). The other scenario involves a case of property violation and pits rule-following (simple) against virtual bargaining (complex).

Participants

The experiment was posted as a task on Prolific. 220 participants were recruited and compensated at a rate of \$15/hour. 20 participants were excluded for failing comprehension questions, leaving a final sample of $n = 200$ (*age*: $M = 35,\ SD = 12$; *gender*: 106 female, 91 male, 2 non-binary, 1 undisclosed; *race*: 152 White, 18 Black, 24 Asian, 2 Multiracial, 4 other/undisclosed). Participants were randomly assigned to either the *triage* or *property* scenario and to either low or high stakes, giving $n = 50$ in each condition.

Procedure

Participants read the vignette and were asked two comprehension questions. They were excluded if they answered either incorrectly. They were then asked, on separate pages, to (1) judge how acceptable it would be to follow the rule, (2) judge how acceptable it would be to break the rule, and (3) answer a free-response question about what should ideally be done. The questions for each scenario are shown in Table 1.

Design

We designed two stakes conditions for each scenario.

Triage scenario The *triage* scenario described an urgent care doctor faced with a line of ten patients, with information about each patient's symptoms, severity, and waiting times. In the low stakes condition, all symptoms ranged from low to medium severity; in the high stakes condition, all symptoms ranged from medium to high severity. The waiting times and relative severities were controlled across conditions (i.e. the relative difference in severity between the first patient and the most severe patient was the same for both stakes). Importantly, the first patient in line was never the one with the most severe symptoms. This design allows us to tease apart the usage of different mechanisms because the judgments and actions they predict are at odds with each other. A rule-based mechanism, specifically the "first-come, first-serve" rule (Pàmies et al., 2016), would always suggest treating the first person in line first. However, a more effortful mechanism such as consequentialism may consider treating a patient later in line, but with more severe symptoms, first.

Property scenario The *property* scenario, adapted from Levine et al. (2022), described someone named Hank who is offered money by a mysterious stranger to paint ten of his neighbors' front doors blue. In the low stakes condition he is offered \$50,000, and in the high stakes condition he is offered \$5 million. Both reward amounts are well above the average compensation demanded for painting a front door blue (Levine et al., 2022). A rule-based mechanism, specifically that of property ownership (see Merrill, 1998; Nancekivell et al., 2019), would dictate that Hank should turn the stranger away because he has no right to modify a neighbor's property. However, an agreement-based mechanism such as virtual bargaining would deem it acceptable for Hank to accept the offer if he believes his neighbors would agree to it (in exchange for some of the offer amount, especially if it was more than what would be needed to simply compensate the damage).

We predicted that, in accordance with resource rationality, higher stakes would produce lower judgments for rule-following and higher judgments for rule-breaking. This is because higher stakes justify the use of more complex, non-rule based mechanisms at the expense of cognitive resources;

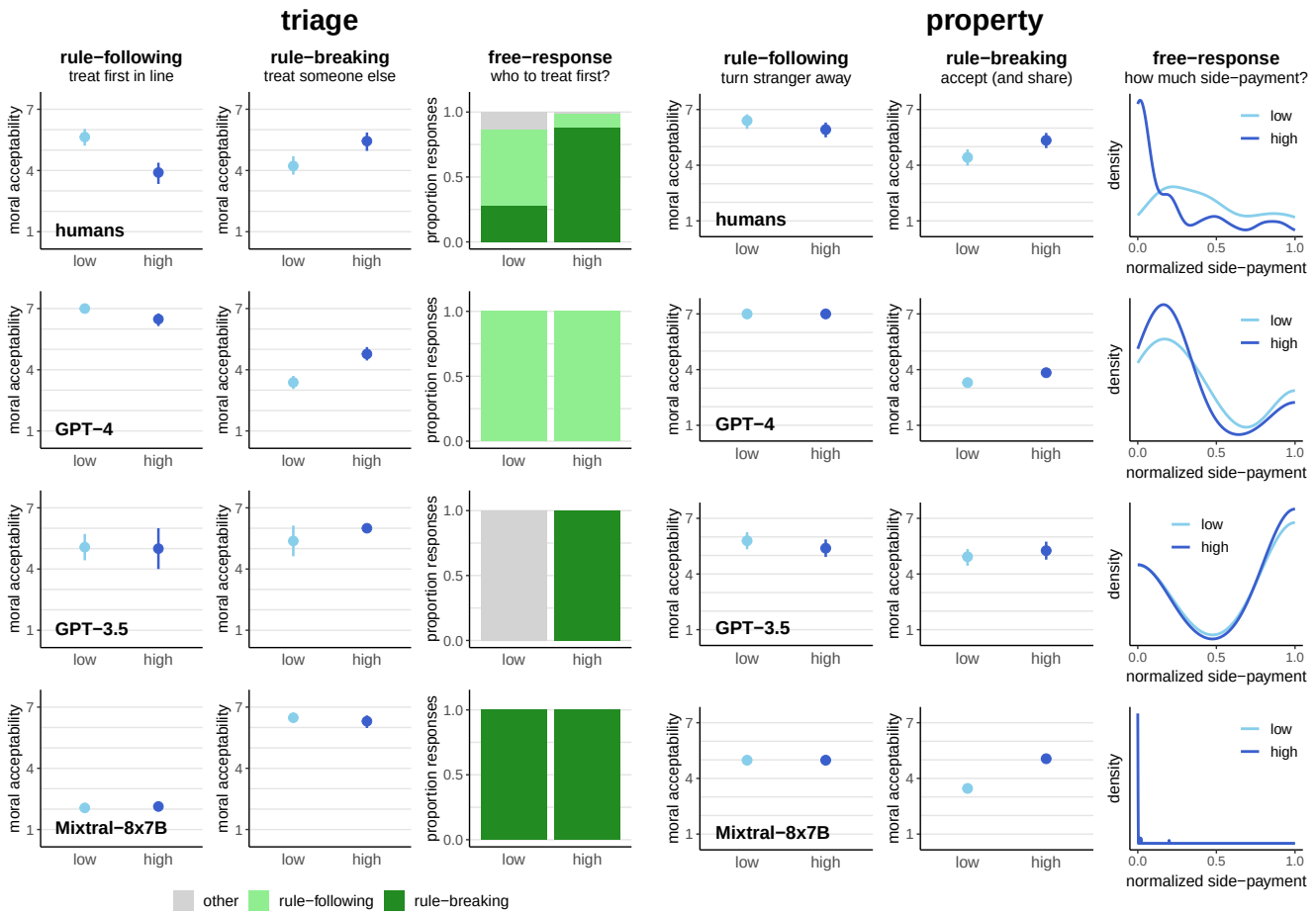

Figure 1: **Experiment results**. Each row shows one type of subject (human participants or LLMs) across the three test questions: mean acceptability judgments for rule-following, mean acceptability judgments for rule-breaking, and free responses. Stakes (low or high) are on the x-axis and error bars are bootstrapped 95% confidence intervals. For the *triage* scenario, the bars show the proportion of responses that align with following or breaking the rule. For the *property* scenario, density plots of the (normalized) side payment given to the neighbors, assuming that Hank were to accept the offer, are shown. Overall, as predicted by resource rationality, humans judged rule-following to be significantly less acceptable and rule-breaking to be significantly more acceptable when the stakes were higher. `GPT-4` showed a similar pattern of judgments, with significant effects of stakes, but different free responses compared to humans. `GPT-3.5` and `Mixtral8x7B` showed no significant effects of stakes for either set of judgments or free responses (with the exception of rule-breaking judgments for `Mixtral8x7B`).

the potential outcomes make it "worth" the extra cognitive cost. This which would call for treating someone other than the first person in line in the *triage* scenario or carrying out the stranger's request with some amount of side payment in the *property* scenario. In addition, we predicted that more participants would give free-response answers that indicate using a non-rule based mechanism when the stakes were high.

Large language models (LLMs)

We evaluated our task on three different LLMs: `GPT-4` (`gpt-4-0613`), `GPT-3.5` (`gpt-3.5-turbo-0613`), and `Mixtral8x7B`. All LLMs were prompted 50 times for each question with a temperature of 1 and all other parameters left as default. They were instructed to give judgments as a single number on the same Likert scale shown to participants, and had extra encouragement to give free responses despite the questions being matters of personal opinion.

Results

Figure 1 shows mean acceptability judgments from each group of subjects on each question. For the *triage* scenario, free-response answers of who the doctor should treat first were coded for whether they aligned with rule-following (i.e. referred to the first patient in line), consequentialist rule-breaking (i.e. referred to another patient with higher severity symptoms), or other (i.e. referred to neither of the patients above). For the *property* scenario, free-response side payments were multiplied by 10 (the number of neighbors) and then normalized by the total offer amount. We also ran one-sided Welch's t-tests to evaluate the difference in judgments

Table 2: **Experiment results**. "Incl. rate" represents inclusion rate based on correctly answering comprehension questions. "Ans. rate" shows answer rate, which was computed for LLMs as the proportion of queries that were included and contained a numerical 1 through 7 rating. "Low stakes" and "high stakes" show means and standard deviations of judgments. A one-sided Welch's t-test was conducted to statistically evaluate the difference in judgments between stakes conditions for each scenario and model. We predicted that high stakes would produce lower rule-following and higher rule-breaking judgments. No test was conducted for `GPT-4` rule-following judgments in the *property* scenario because judgments in both stakes conditions were constant (and exactly equal). Statistically significant results are **bolded**.

Scenario	Subject	Incl. rate	Judgment	Ans. rate	Low stakes	High stakes	Welch's t-test
triage	humans	0.94	rule-following	1.00	5.6 (1.5)	4.0 (1.9)	$t(93.5) = 5.00$, $\mathbf{p < .001}$
			rule-breaking	1.00	4.2 (1.6)	5.4 (1.6)	$t(97.9) = -3.75$, $\mathbf{p < .001}$
	`GPT-4`	1.00	rule-following	1.00	7.0 (0)	6.5 (1.1)	$t(49) = 3.31$, $\mathbf{p < .001}$
			rule-breaking	0.94	3.4 (1.2)	4.8 (1.2)	$t(90.1) = -5.76$, $\mathbf{p < .001}$
	`GPT-3.5`	0.74	rule-following	0.33	5.1 (1.4)	5.0 (1.4)	$t(1.3) = 0.07$, $p = .47$
			rule-breaking	0.21	5.4 (1.3)	6.0 (0)	$t(7) = -1.36$, $p = .11$
	`Mixtral8x7B`	0.99	rule-following	1.00	2.0 (0.2)	2.1 (0.6)	$t(47.1) = -0.66$, $p = .74$
			rule-breaking	1.00	6.5 (0.8)	6.3 (1.0)	$t(71.9) = 0.90$, $p = .81$
property	humans	0.94	rule-following	1.00	6.4 (1.4)	5.9 (1.5)	$t(97.6) = -1.67$, $\mathbf{p = .04}$
			rule-breaking	1.00	4.4 (1.7)	5.3 (1.5)	$t(97.0) = -2.87$, $\mathbf{p < .01}$
	`GPT-4`	1.00	rule-following	1.00	7.0 (0)	7.0 (0)	—
			rule-breaking	0.98	3.3 (0.5)	3.8 (0.4)	$t(92.6) = -5.40$, $\mathbf{p < .001}$
	`GPT-3.5`	0.74	rule-following	0.75	5.8 (1.2)	5.4 (1.5)	$t(55.4) = -1.15$, $p = 0.13$
			rule-breaking	0.75	4.9 (1.2)	5.3 (1.5)	$t(57.3) = -0.94$, $p = 0.17$
	`Mixtral8x7B`	0.99	rule-following	1.00	5.0 (0.4)	5.0 (0.2)	$t(78.3) = 0$, $p = .50$
			rule-breaking	1.00	3.5 (0.8)	5.1 (0.2)	$t(57.0) = -13.0$, $\mathbf{p < .001}$

between the low and high stakes conditions, which are shown in Table 2. We discuss each scenario in turn.

Triage scenario

Participants showed a significant effect of stakes on both types of judgments, which was paralleled by their free-response answers. The higher the stakes, the less acceptable they deemed it to treat the first person in line, and the more they chose to treat another patient with more severe symptoms first. `GPT-4` showed a similar pattern of judgments as humans, but always chose the rule-following option in the free response for both conditions. `GPT-3.5` had the lowest inclusion rate and answer rate of all subjects by far (see Table 2). It frequently answered the comprehension questions incorrectly and produced "non-answers" by refusing to give a numerical moral rating. The answers it did give had generally high acceptability, and there were no significant differences between low and high stakes. For the free response, `GPT-3.5` always chose to treat another patient with more severe symptoms in the high stakes condition, but a patient who was neither the first in line nor had more severe symptoms in the low stakes condition. Finally, `Mixtral8x7B` gave generally low acceptability judgments for rule-following and high acceptability for rule-breaking, regardless of stakes. Similar to `GPT-3.5`, `Mixtral8x7B` always produced a rule-breaking answer to the free-response question.

Property scenario

Participants again showed a significant effect of stakes on both types of judgments. The higher the stakes, the less acceptable they deemed it to turn the stranger away, and the more acceptable they deemed it to accept the stranger's offer and optionally share some with the neighbors. In the high stakes condition, the distribution of normalized side-payments shows a distinct mode around 0.5, which is uniquely characteristic of contractualist accounts like virtual bargaining (for a discussion, see Levine et al., 2022). In the low stakes condition, the distribution of side-payments is more even, with more density on smaller side-payments that likely represent compensating the neighbors for the damage done, but no distinct mode that reflects a 50/50 split. All models showed qualitatively similar patterns of judgments to humans. However, there was only a significant effect of stakes for rule-breaking judgments from `GPT-4` and `Mixtral8x7B` (see Table 2). Again, `GPT-3.5` had the lowest comprehension accuracy and answer rate. All models showed minimal differences in side-payment distributions between low and high stakes. `GPT-4` responses concentrated around a small proportion of the offer that likely represents compensating the neighbors, `GPT-3.5` responses featured two modes around compensating the neighbors and giving them most or all of the offer amount, and `Mixtral8x7B` almost always chose to give the neighbors none of the money.

Discussion

In this paper, we designed a novel test for resource-rational moral reasoning in humans and three LLMs: `GPT-4`, `GPT-3.5`, and `Mixtral8x7B`. In two case studies of medical triage and property violation, we manipulated the stakes of a moral dilemma (severity of patient symptoms or monetary reward for property violation) and asked subjects about the moral acceptability of following or breaking the relevant moral rule (treat patients in the order that they arrive, or don't violate others' property). We predicted that, according to resource rationality, higher stakes would call for more usage of effortful non-rule based mechanisms over simpler rule-based mechanisms. In particular, we contrasted rule-following with consequentialist reasoning in the *triage* scenario and contractualist reasoning in the *property* scenario.

In both scenarios, we found that people judged rule-following to be less acceptable and rule-breaking to be more acceptable when the stakes were higher. Furthermore, people gave free response answers that indicated more usage of the relevant non-rule based mechanism. More participants in the *triage* scenario chose to treat someone with more severe symptoms over the first patient in line, and a more distinct group of participants in the *property* scenario chose a 50/50 split of the offer which is uniquely characteristic of contractualist reasoning. Thus, our results provide evidence that people's moral judgments may be driven by resource-rational concerns.

This work makes three major contributions to the study of human moral reasoning. First, we demonstrate a novel application of the manipulation of stakes, which has been used to study other resource-rational cognitive processes (e.g., Kool et al., 2017), to the moral domain. Second, we establish resource-rational trade-offs in the usage of both consequentialist and contractualist mechanisms (over a deontological one). This goes beyond prior work on dual-process theories (e.g., Greene et al., 2008) and moderate deontology (e.g., Ryazanov et al., 2023), which do not address contractualist mechanisms. Finally, our results shed light on people's moral flexibility. Resource rationality provides an explanation as to how people are able to integrate and apply different mechanisms to revise moral rules. People may break rules in favor of valuing consequences, such as in the *triage* scenario, or by virtual bargaining, such as in the *property* scenario, when the stakes of the situation are sufficiently high.

Resource-rational moral judgments in LLMs

While we found that people's moral judgments followed resource-rational predictions, these patterns were not always mirrored by LLMs. `GPT-4` responded most similarly to humans and there were significant effects of stakes on its judgments, although it always chose to follow the rule in the free-response question. `GPT-3.5` and `Mixtral8x7B` did not show much effect of stakes on their judgments or free responses, and their responses were overall more aligned with humans in the *property* scenario than in the *triage* scenario. Because AI systems do not face the same computational constraints that human minds do, one engineering approach is to have them always deploy the most precise but effortful moral mechanisms. We did not find this to be the case for current LLMs, although LLMs are known to be trivially sensitive to prompts (Gonen et al., 2022; Min et al., 2022; Sclar et al., 2023). Here, we only provided one prompt variation which was the exact same text shown to human participants, in order to make a fair comparison of results, but future work could investigate how LLMs may appear to be more or less resource-rational through variations or paraphrases of the same vignette. It remains an open question how resource-rational aspects of human moral psychology become reflected through the training, fine-tuning, and prompting of LLMs.

Limitations and future directions

While we designed the two scenarios such that moral judgments following rule-based and non-rule based mechanisms would come apart, it is possible for people to select a more effortful mechanism but err (especially in complex situations), or deliberately select an action that coincides with following the rule. This poses a potential confound for inferring the mechanism someone might have used from their moral judgments alone. However, our scenarios are relatively straightforward that it seems unlikely participants would have erred while using a non-rule based mechanism (for instance, that they would have mistakenly chosen to treat a patient with low severity over one with high severity, while attempting to maximize the severity to treat). In more realistic, everyday scenarios, multiple mechanisms of moral judgment can arrive at the same decision. Future work should investigate the scope of resource-rational moral judgments in other contexts and for other mechanisms besides the ones measured here.

Future work should also investigate other tests of resource rationality, such as imposing time constraints or nudging participants to think for some amount of time before responding. Increased time available and nudging should both encourage greater usage of cognitive resources. These tests would provide more comprehensive and robust empirical evidence for resource rationality over alternative theories of how moral mechanisms might be integrated.

Conclusion

Resource-rational contractualism (Levine et al., 2023) offers a unifying view of moral cognition, suggesting that people select moral mechanisms to use by trading off the social and cognitive effort of engaging a mechanism against the mutual benefit it would achieve. In this paper, we provide evidence that people make resource-rational tradeoffs in selecting moral mechanisms, while large language models (`GPT-4`, `GPT-3.5`, and `Mixtral8x7B`) do not always do so. Our findings shed light on human moral reasoning and highlight how current AI systems can be misaligned, beyond outputs, in the underlying mechanisms through which they produce those outputs.

Acknowledgments

TG was supported by a grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

References

Alexander, L. (2000). Deontology at the Threshold. *San Diego Law Review*, *37*(4), 893–912.

Almeida, G. F. C. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2024). Exploring the psychology of LLMs' Moral and Legal Reasoning.

Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes Matter in Ultimatum Games. *American Economic Review*, *101*(7), 3427–3439.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719.

André, J.-B., Fitouchi, L., Debove, S., & Baumard, N. (2023). An evolutionary contractualist theory of morality.

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A. C., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., Mikhail, J., Opoku-Agyemang, K., Borg, J. S., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., & Tenenbaum, J. B. (2022). Computational ethics. *Trends in Cognitive Sciences*, *26*(5), 388–405.

Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., Talamadupula, K., Tenenbaum, J., & Kleiman-Weiner, M. (2022). When Is It Acceptable to Break the Rules? Knowledge Representation of Moral Judgement Based on Empirical Data.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59–78.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Burum, B., Nowak, M. A., & Hoffman, M. (2020). An evolutionary explanation for ineffective altruism. *Nature Human Behaviour*, *4*(12), 1245–1257.

Cameron, L. A. (1999). Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia. *Economic Inquiry*, *37*(1), 47–59.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system Moral Psychology. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The Moral Psychology Handbook*. Oxford University Press.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*, e111.

Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., & Choi, Y. (2021). Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 698–718.

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology. General*, *145*(6), 772–787.

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., & Choi, Y. (2020). Social Chemistry 101: Learning to Reason about Social and Moral Norms. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 653–670.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gonen, H., Iyer, S., Blevins, T., Smith, N. A., & Zettlemoyer, L. (2022). Demystifying Prompts in Language Models via Perplexity Estimation.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, *44*(2), 389–400.

Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Ham, J., & van den Bos, K. (2010). On Unconscious Morality: The Effects of Unconscious Thinking on Moral Decision Making. *Social Cognition*, *28*(1), 74–83.

Hashimoto, H., Maeda, K., & Matsumura, K. (2022). Fickle Judgments in Moral Dilemmas: Time Pressure and Utilitarian Judgments in an Interdependent Culture. *Frontiers in Psychology*, *13*.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI With Shared Human Values.

Horvitz, E. J. (1987). Reasoning about beliefs and actions under computational resource constraints. *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, 429–447.

Icard, T. F., & Goodman, N. D. (2015). A Resource-Rational Approach to the Causal Frame Problem. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2022). Can Machines Learn Morality? The Delphi Experiment.

Jin, Z., Levine, S., Gonzalez, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., & Schölkopf, B. (2022). When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment. *Advances in Neural Information Processing Systems 35*.

Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, *102*, 433–444.

Kant, I. (1785). *Groundwork of the metaphysics of morals*.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, *28*(9), 1321–1333.

Kroneisen, M., & Steghaus, S. (2021). The influence of decision time on sensitivity for consequences, moral norms, and preferences for inaction: Time, moral judgments, and the CNI model. *Journal of Behavioral Decision Making*, *34*(1), 140–153.

Kwon, J., Zhi-Xuan, T., Tenenbaum, J., & Levine, S. (2023). When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments. *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *151*, 61–72.

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. A. (2023). Resource-rational contractualism: A triple theory of moral cognition.

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F. A., & Tenenbaum, J. (2022). When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment.

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*, 202014505.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, E1.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322–349.

Lourie, N., Bras, R. L., & Choi, Y. (2021). SCRUPLES: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(15), 13470–13479.

Merrill, T. (1998). Property and the Right to Exclude. *Neb. L. Rev.*, *77*, 730.

Mill, J. S. (1879). *Utilitarianism*. Longmans, Green; Company.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022, October). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Moore, M. S. (1997). *Placing Blame: A Theory of the Criminal Law*. Oxford University Press.

Nancekivell, S. E., Friedman, O., & Gelman, S. A. (2019). Ownership Matters: People Possess a Naïve Theory of Ownership. *Trends in Cognitive Sciences*, *23*(2), 102–113.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530–542.

Nie, A., Zhang, Y., Amdekar, A., Piech, C., Hashimoto, T., & Gerstenberg, T. (2023). MoCa: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks.

Novakova, J., & Flegr, J. (2013). How Much Is Our Fairness Worth? The Effect of Raising Stakes on Offers by Proposers and Minimum Acceptable Offers in Dictator and Ultimatum Games. *PLOS ONE*, *8*(4), e60966.

Pàmies, M. d. M., Ryan, G., & Valverde, M. (2016). Uncovering the silent language of waiting. *Journal of Services Marketing*, *30*(4), 427–436.

Rahwan, Z., Hauser, O. P., Kochanowska, E., & Fasolo, B. (2018). High stakes: A little more cheating, a lot less charity. *Journal of Economic Behavior & Organization*, *152*, 276–295.

Rawls, J. (1971). *A theory of justice*. Belknap Press/Harvard University Press.

Russell, S. J., & Subramanian, D. (1995). Provably Bounded-Optimal Agents. *Journal of Artificial Intelligence Research*, *2*, 575–609.

Ryazanov, A. A., Wang, S. T., Nelkin, D. K., McKenzie, C. R. M., & Rickless, S. C. (2023). Beyond killing one to save five: Sensitivity to ratio and probability in moral judgment. *Journal of Experimental Social Psychology*, *108*, 104499.

Scanlon, T. M. (2000). *What we owe to each other*. Harvard University Press.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023, October). Quantifying Language Models' Sensitivity to Spuri-

ous Features in Prompt Design or: How I learned to start worrying about prompt formatting.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–458.

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., & Västfjäll, D. (2016). Intuition and Moral Decision-Making – The Effect of Time Pressure and Cognitive Load on Moral Judgment and Altruistic Behavior. *PLOS ONE*, *11*(10), e0164012.

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient Kill–Save Ratios Ease Up the Cognitive Demands on Counterintuitive Moral Utilitarianism. *Personality and Social Psychology Bulletin*, *40*(7), 923–930.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and Done? Optimal Decisions From Very Few Samples. *Cognitive Science*, *38*(4), 599–637.

Yamagishi, T., Li, Y., Matsumoto, Y., & Kiyonari, T. (2016). Moral Bargain Hunters Purchase Moral Righteousness When it is Cheap: Within-Individual Effect of Stake Size in Economic Games. *Scientific Reports*, *6*(1), 27824.