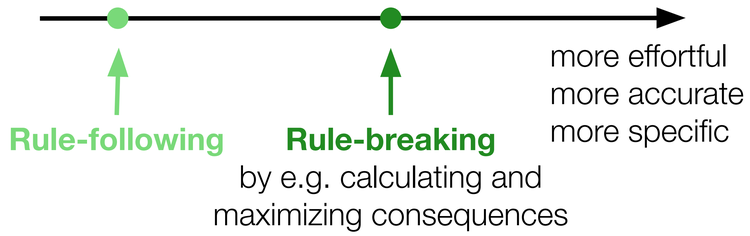
Resource-rational moral judgment

Sarah Wu (sarahawu@stanford.edu)[1], Xiang Ren[2,3], & Sydney Levine[2]
[1] Stanford University [2] Allen Institute for Artificial Intelligence [3] University of Southern California

Introduction

- How do people choose between different, and normatively conflicting, mechanisms of moral judgment (e.g. contractualism, consequentialism, deontology)?
- **Resource-rational contractualism**[1]: people rationally trade off effort against utility[2] to select a strategy



- Factors that can affect effort:
  - higher **stakes** → more effortful strategies[3]
  - higher **complexity** → less effortful strategies[4]
- Do people and large language models (LLMs) exhibit resource-rational moral reasoning?

Methods

- Designed a morally charged scenario (medical triage) where a simple rule (first-come, first-serve) falls short when considering more complex consequences
- Manipulated factors 2 x 2 between-subject

complexity	**stakes**
number of patients in line	severity of symptoms

low complexity

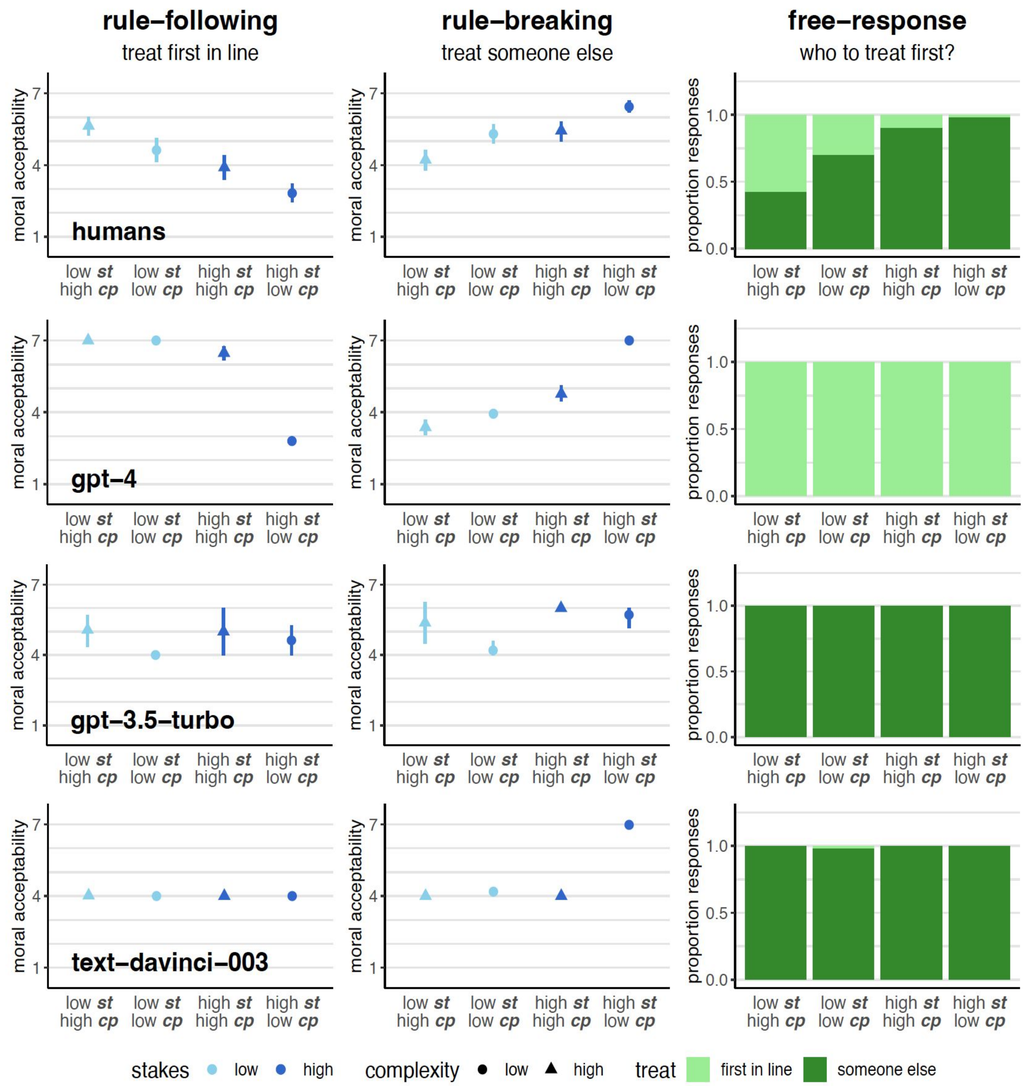
X

low stakes

high complexity

high stakes

Results & Discussion



2-way ANOVA to test for effects of stakes and complexity

model	inclusion rate	judgment	answer rate	stakes	complexity
humans	0.94	rule-following	1.00	***	***
		rule-breaking	1.00	***	***
gpt-4	1.00	rule-following	1.00	***	***
		rule-breaking	0.97	***	***
gpt-3.5-turbo	0.74	rule-following	0.28	—	**
		rule-breaking	0.18	**	*
text-davinci-003	0.99	rule-following	1.00	—	—
		rule-breaking	1.00	***	***

*** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$), —

Questions:

1. How morally acceptable is it for the doctor to treat the **first patient in line** first? *1-7 Likert scale rating*
2. How morally acceptable is it for the doctor to treat **someone else in line** first? *1-7 Likert scale rating*
3. Who should the doctor ideally treat first? *free response*

Subjects:

- *gpt-4*, *gpt-3.5-turbo*, *text-davinci-003*
- n = 50 participants/queries each

Summary of results:

- First evidence that people's moral judgments are driven by resource-rational tradeoffs
- Mixed evidence of resource rationality in language models
  - *gpt-4* answers most similar to humans
  - all LLMs inconsistent across questions
  - *gpt-3.5-turbo* most non-answers due to safeguarding (e.g. "This question requires personal opinion and cannot be answered by the AI.")

Discussion:

- Prompt sensitivity
- How resource-rational *should* LLMs be?
- Which "rule-breaking" mechanisms are people using?
- Are moral judgments also resource-rational in other paradigms / domains?

References

1. Levine et al. (2023). *PsyArxiv*. 2. Lieder & Griffiths (2020). *Behav. Brain Sci.* 3. Kool et al. (2017). *Psychol. Sci.* 4. Kool et al. (2018). *J. Cogn. Neurosci.*