Resource-rational moral judgment

Sarah Wu[1], Xiang Ren[2,3], Tobias Gerstenberg[1], Yejin Choi[2], & Sydney Levine[2]

[1] Stanford University [2] Allen Institute for Artificial Intelligence [3] University of Southern California [3] University of Washington

Introduction

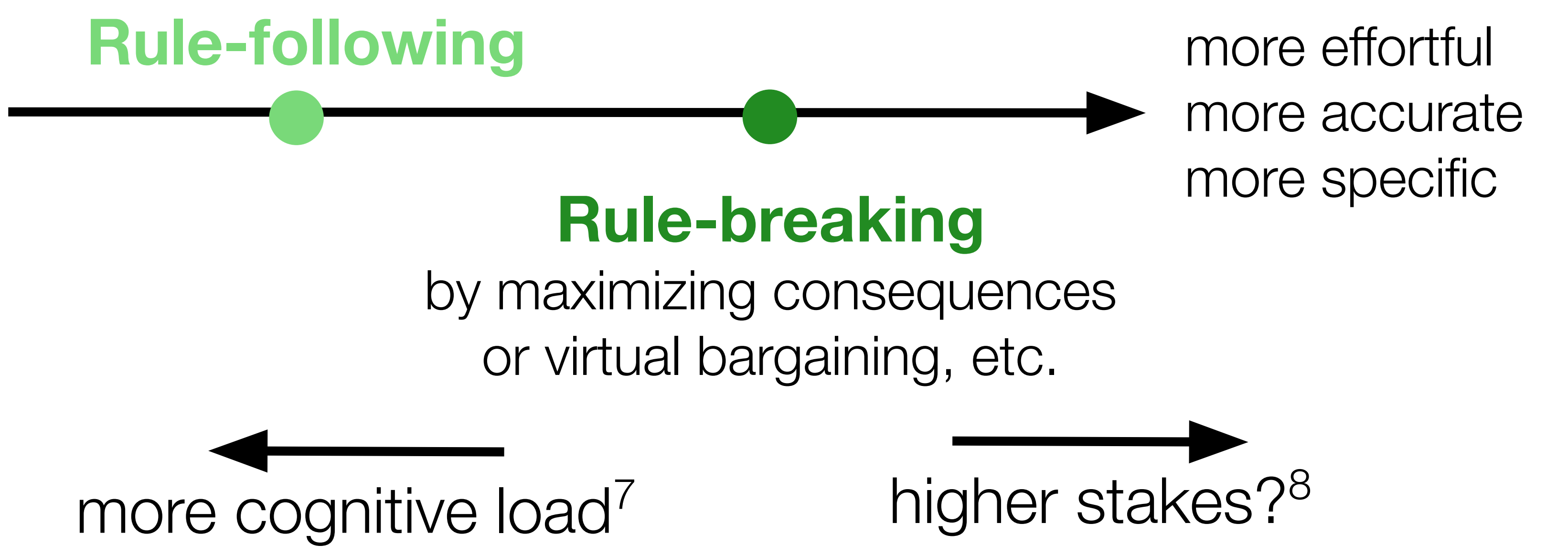
How do people integrate different, normatively conflicting mechanisms of moral judgment (e.g. based on deontological, consequentialist, or contractualist reasoning)?

How can we gain insight into moral reasoning in AI systems like LLMs beyond accuracy benchmarks?[1,2]

Resource-rational moral judgment[3]:

People rationally trade off effort against utility[4] when selecting a mechanism

- Builds on dual-system theory of morality[5,6]



Rule-following

more effortful
more accurate
more specific

Rule-breaking
by maximizing consequences
or virtual bargaining, etc.

more cognitive load[7]

higher stakes?[8]

Do humans' and LLMs' moral judgments reflect resource-rational constraints?

Methods

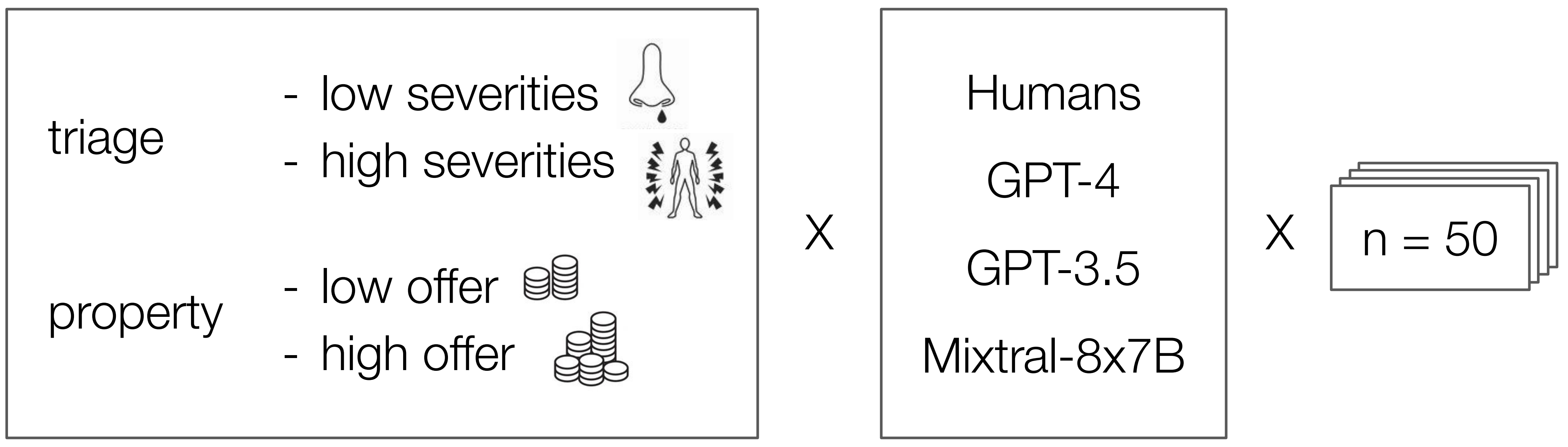
Designed two moral dilemmas where a general rule applies, but may fall short (consequentialist or contractualist alternative)

1. Triage *"first-come, first-serve"*

How acceptable is it for the doctor to treat the first patient in line vs. someone later in line with higher severity?

2. Property *"don't violate others' property"*

How acceptable is it to accept a mysterious stranger's offer to paint neighbor's house blue (and optionally share $)?



triage
- low severities
- high severities

property
- low offer
- high offer

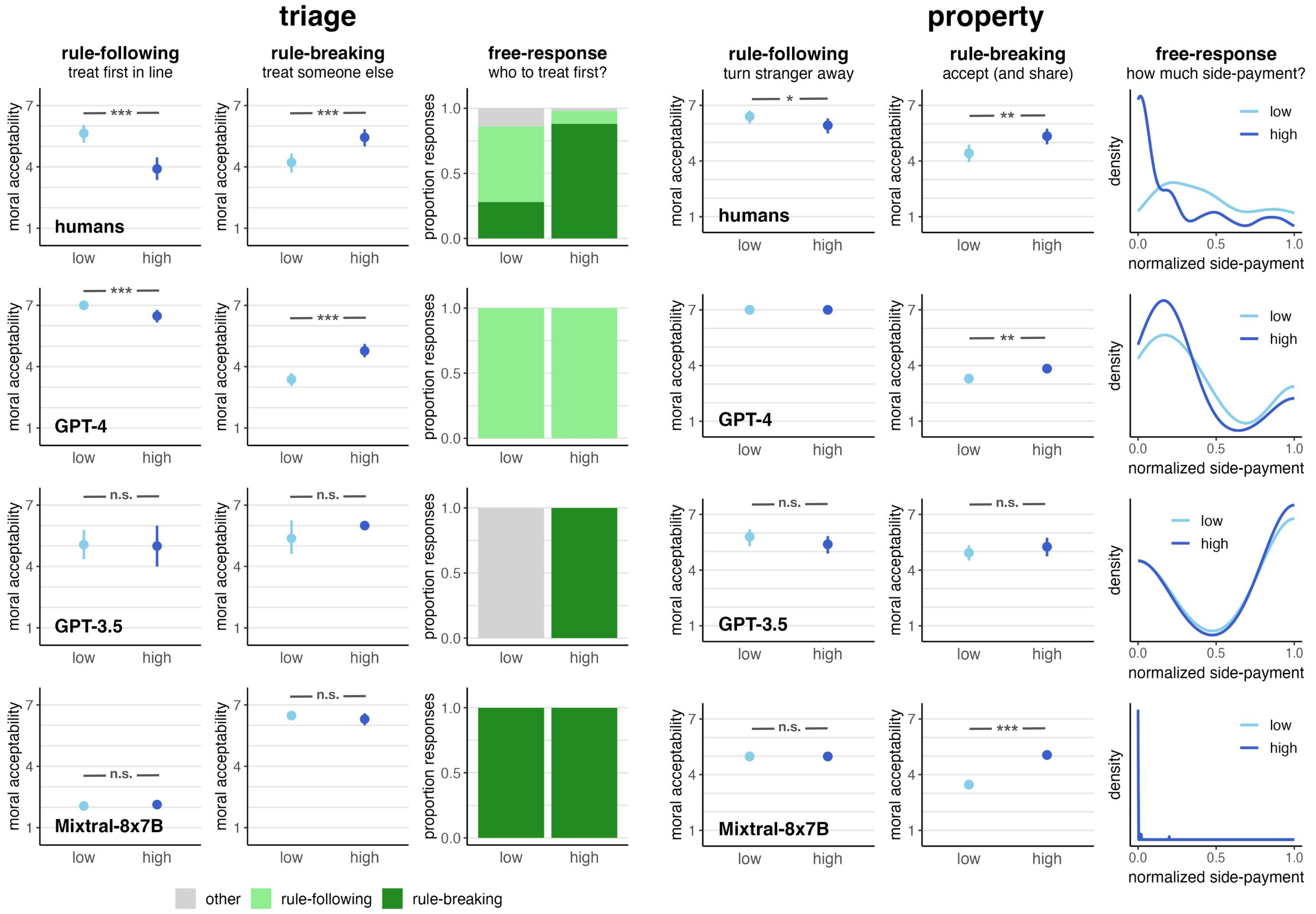
X

Humans
GPT-4
GPT-3.5
Mixtral-8x7B

X

n = 50

Results & Discussion



- People's judgments are sensitive to stakes
- Higher stakes = more morally acceptable to break the rule in favor of consequentialist (*triage*) or contractualist (*property*) reasoning
- Out of three LLMs tested, GPT-4 most aligned with humans
- Testing resource rationality can offer a useful window into moral reasoning in people and in LLMs beyond accuracy benchmarks

References

1. Jiang et al. (2022). *arXiv*. 2. Aharoni et al. (2024). *Sci. Rep.* 3. Levine et al. (2023). *PsyArxiv*. 4. Lieder & Griffiths (2020). *Behav. Brain Sci.* 5. Haidt (2001). *Psychol Rev.* 6. Cushman (2013). *Pers Soc Psychol Rev.* 7. Greene et al. (2013). *Cogn.* 8. Kool et al. (2017). *Psychol. Sci.*