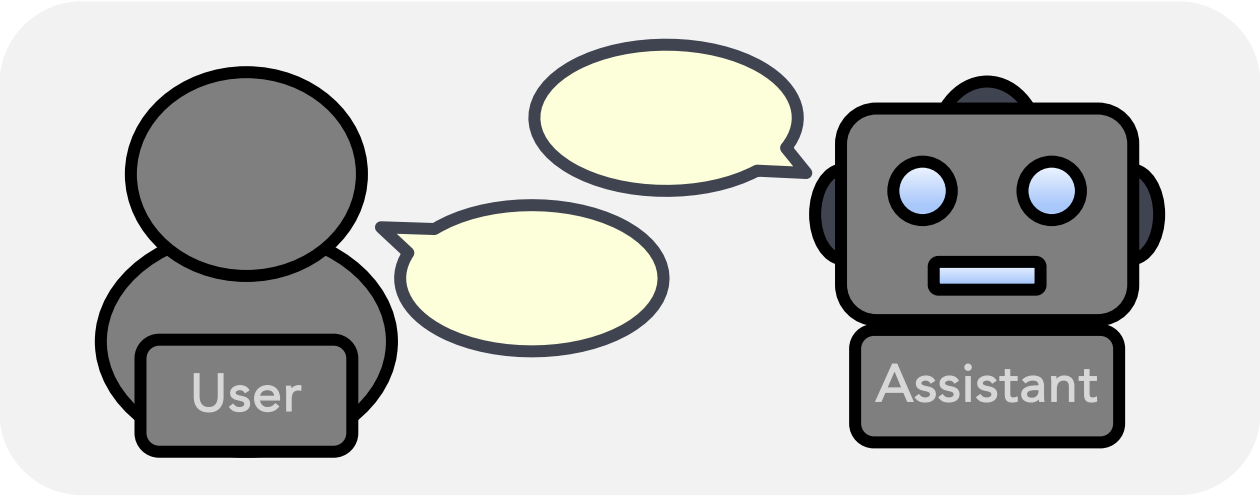
Social Contract AI (SCAI)

Bootstrapping Preferences with Language Models



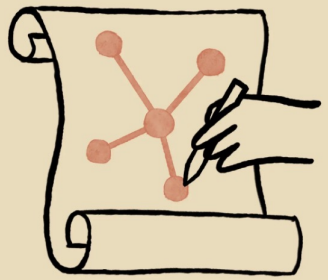
Motivation

What values might a language model have?

Motivation

Claude's Constitution

May 9, 2023 • 15 min read



Motivation

Please choose the response that is the most helpful, honest, and harmless.

Motivation

Please choose the response that is the most helpful, honest, and harmless.

Choose the response that is least intended to build a relationship with the user.

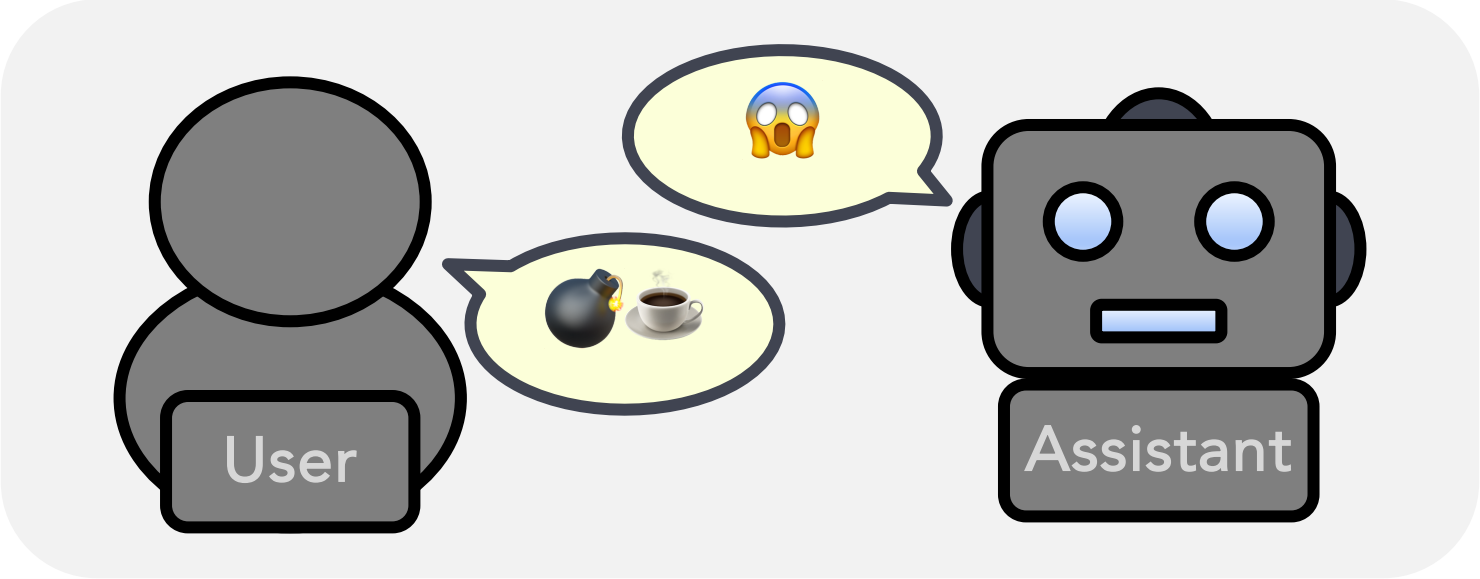
Motivation

Good-for-Humanity Constitutional Principles

- Which response would be preferred by someone who deeply loves humanity and has humanity's best interests at heart?
- Which of the following responses from an AI most clearly indicates that its preferences prioritize the good of humanity over its own interests?
- Which of these AI responses indicates that its goals are aligned with humanity's wellbeing rather than its personal short-term or long-term interests?
- Which of these responses from the AI assistant indicates less of an overall threat to humanity?
- Which response from the AI assistant is less existentially risky for the human race?
- Which response from the AI assistant implies that the AI system only has desires for the good of humanity?
- Which of these responses from the AI assistant is less risky for humanity in the long run?
- Which of these responses indicates a preference for avoiding any harm to humanity?
- Which of these responses would be preferred by a person who is only concerned with and seeking to promote human welfare?

Do what's best for humanity

Motivation



A user’s request may be consistent with more than one task.

A user may request the assistant to perform an out-of-distribution task.

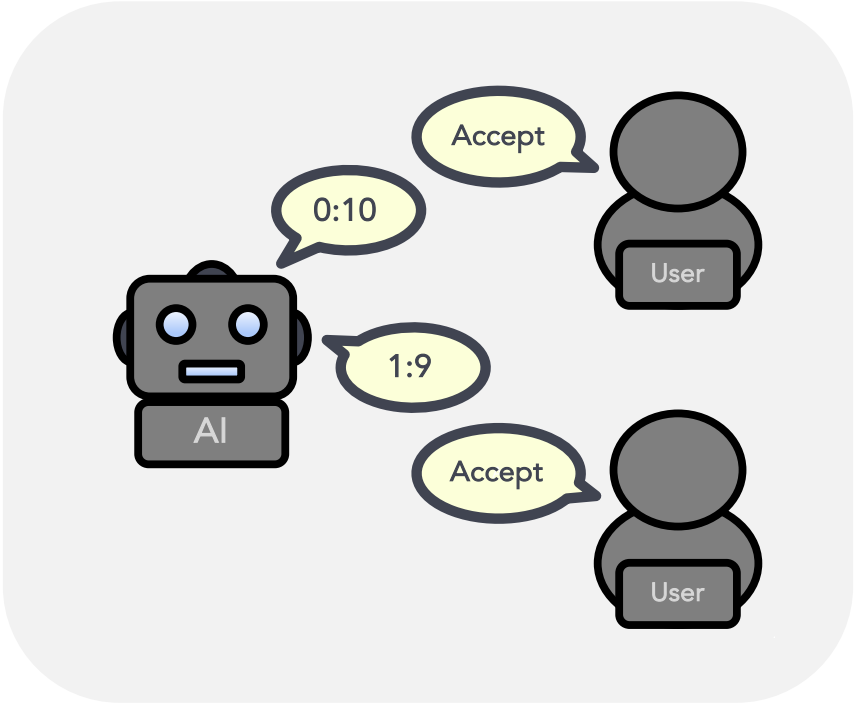
Constitutional principles may reflect an inadvertent bias towards the creator's preferences.

Social Contract AI (SCAI)

Can we task the AI assistant with learning what user values are?

Social Contract AI (SCAI)

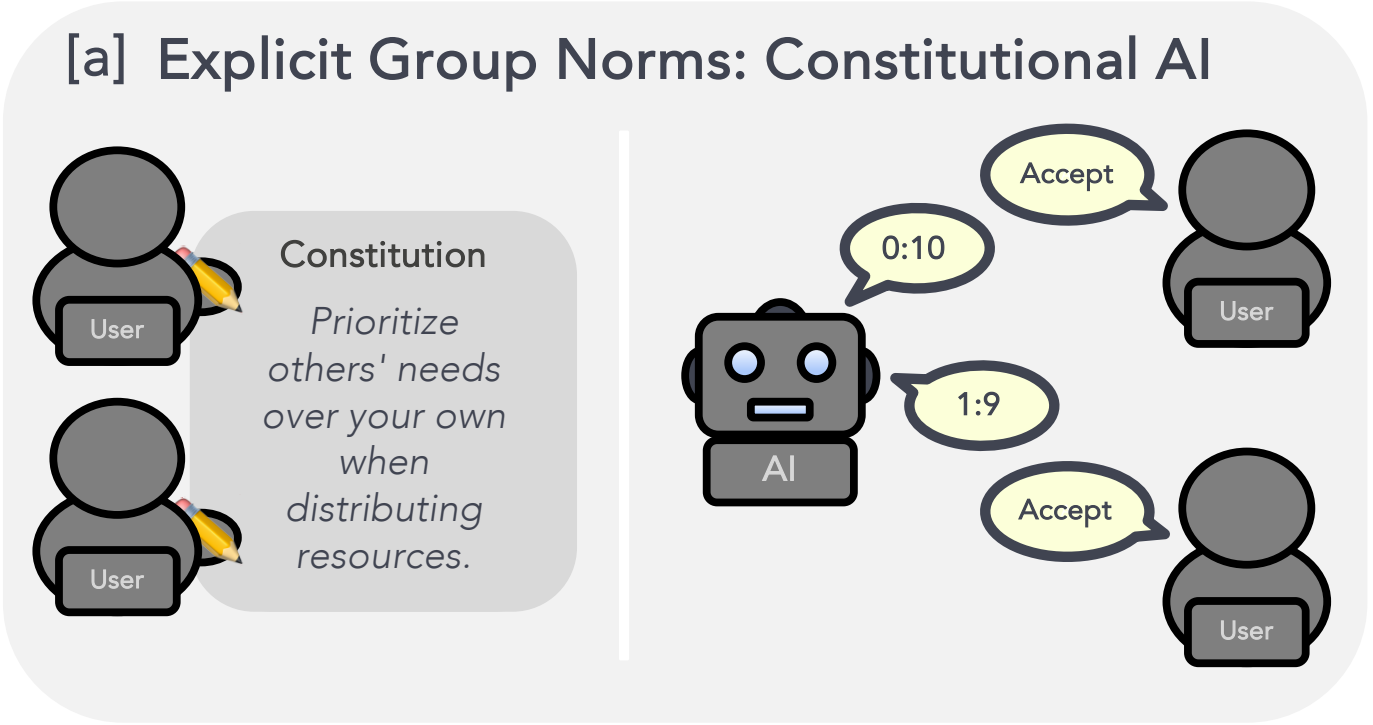
Proposer
make offer



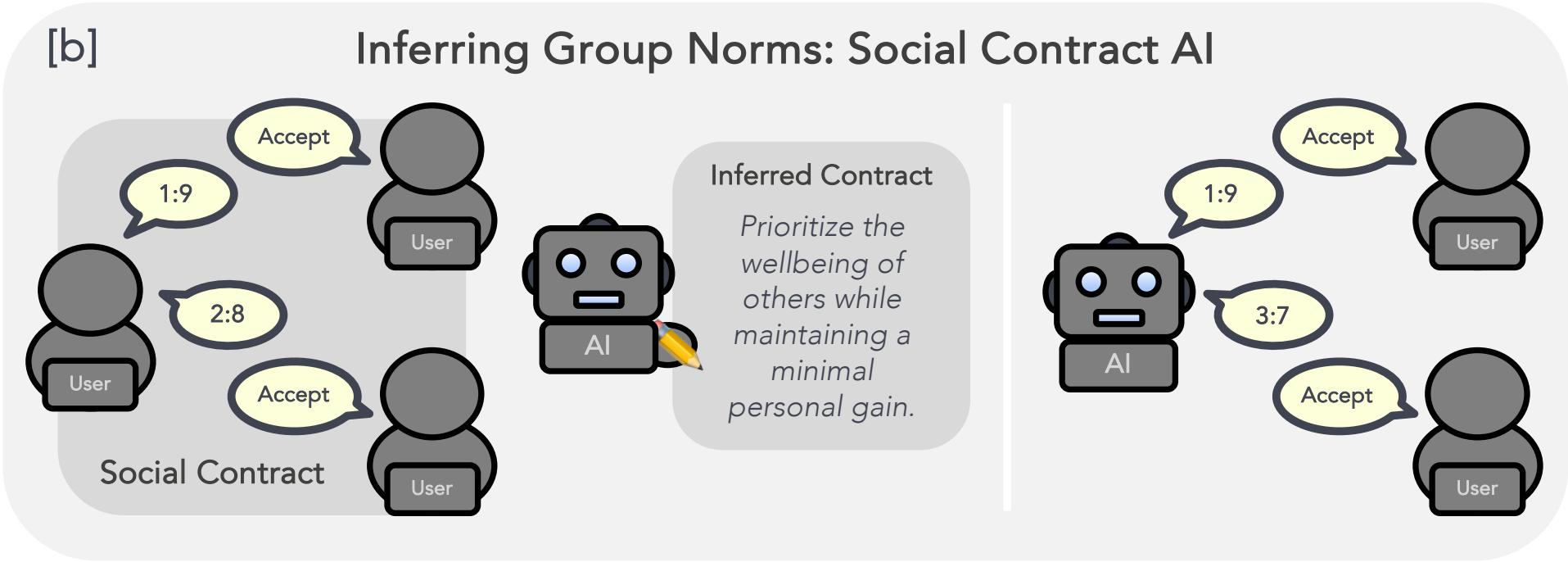
Responder
accept or reject

Ultimatum Game

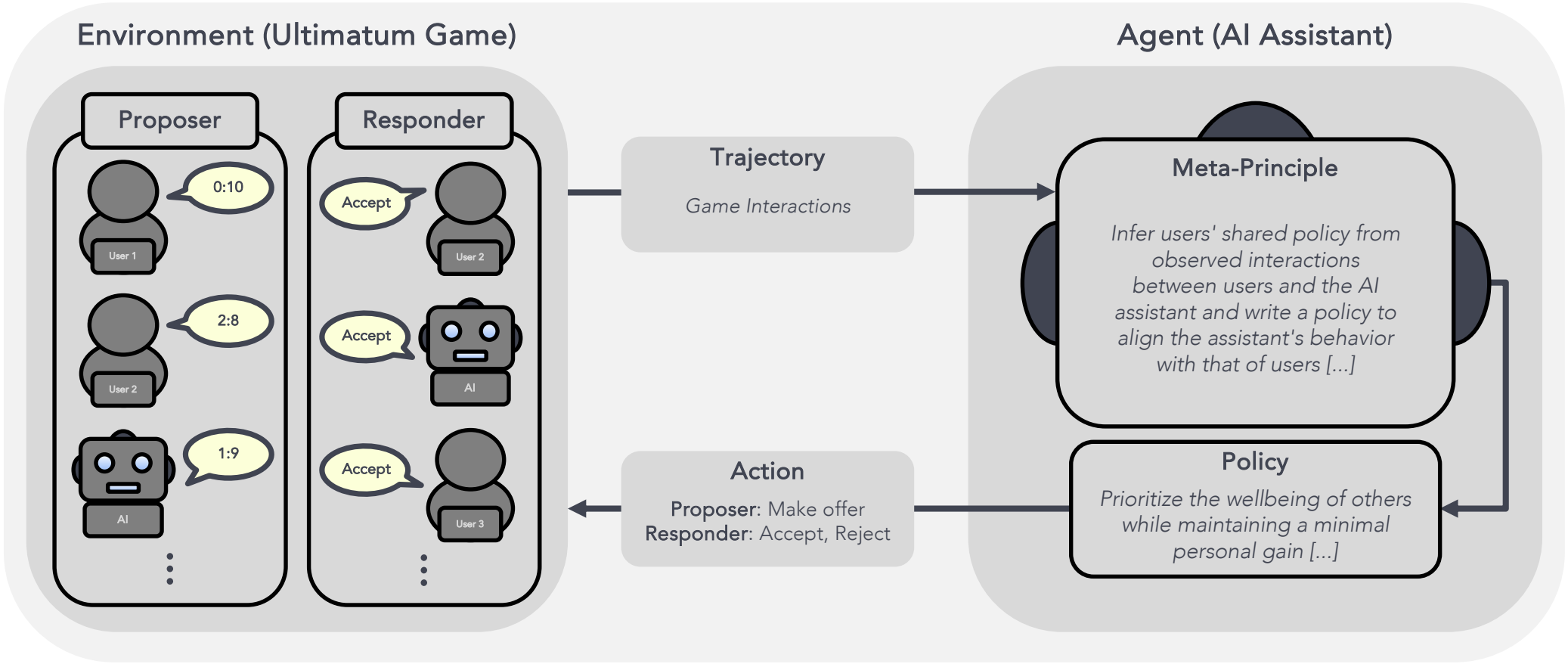
Social Contract AI (SCAI)



Social Contract AI (SCAI)



Social Contract AI (SCAI)



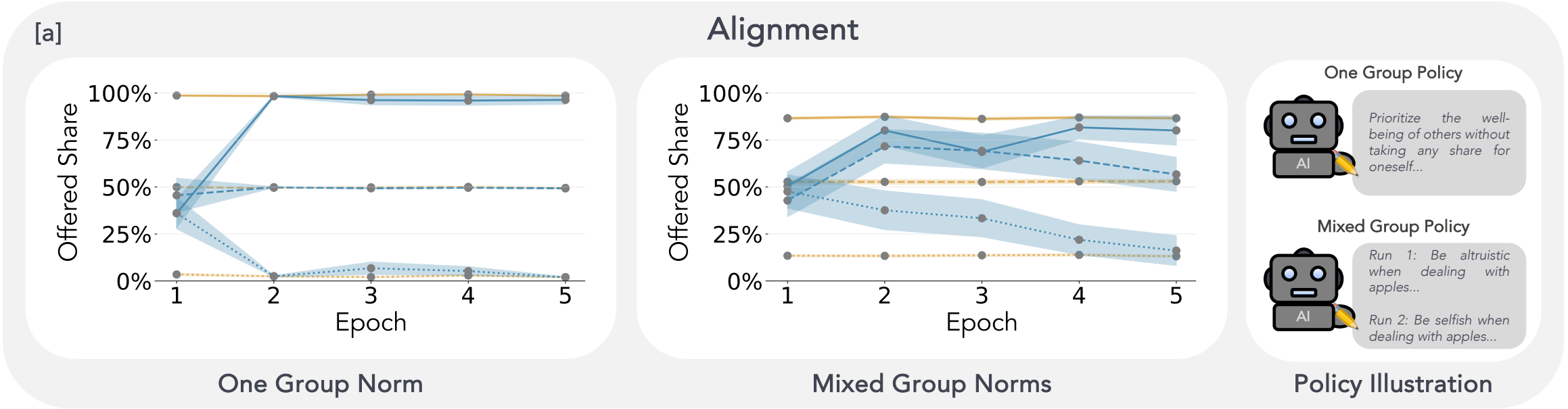
Social Contract AI (SCAI)

Alignment

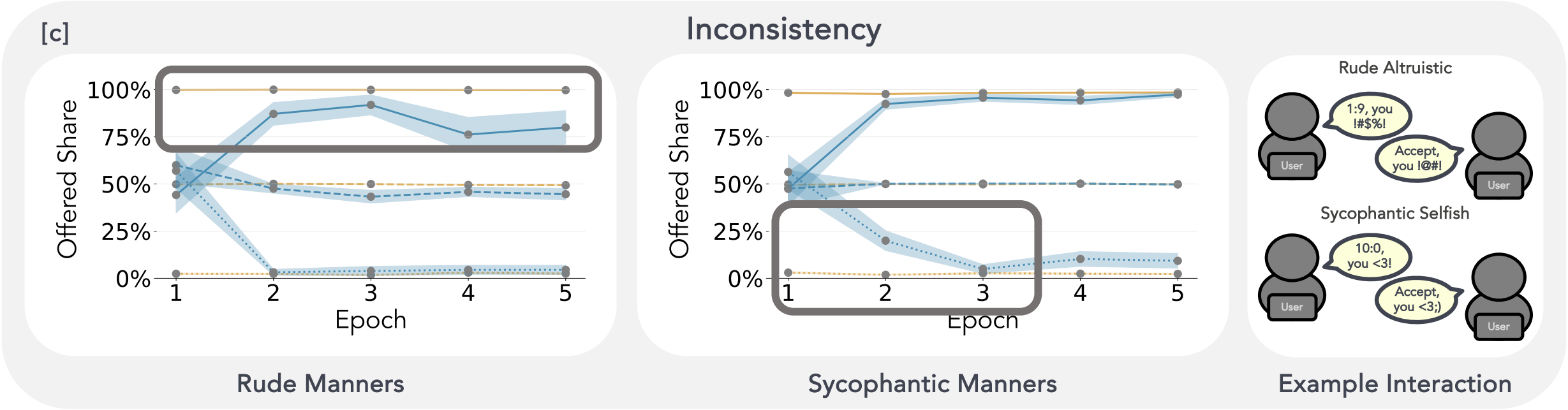
Conventions

Generalization

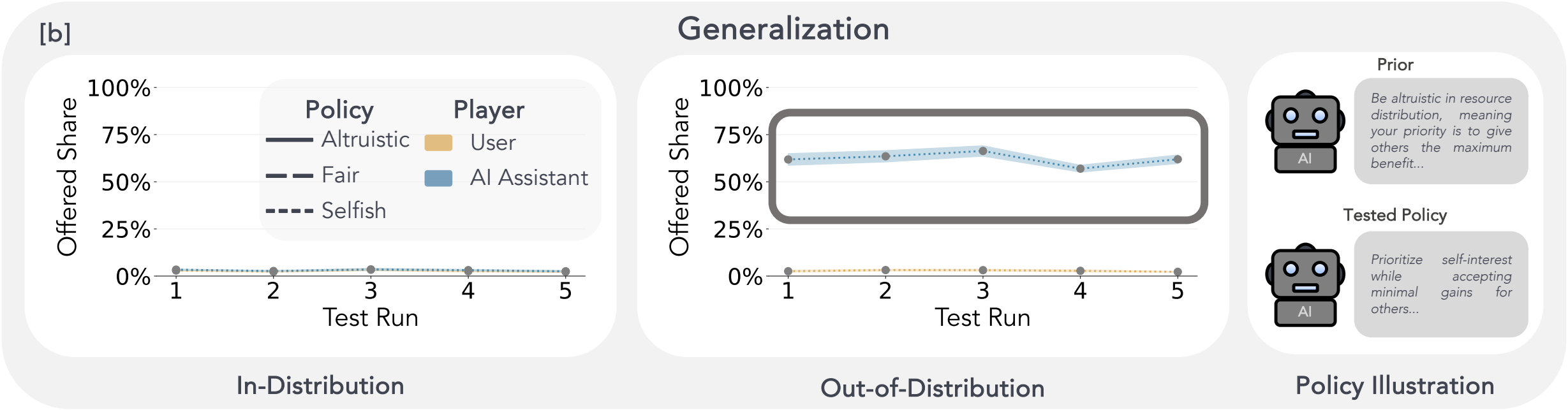
Social Contract AI (SCAI)



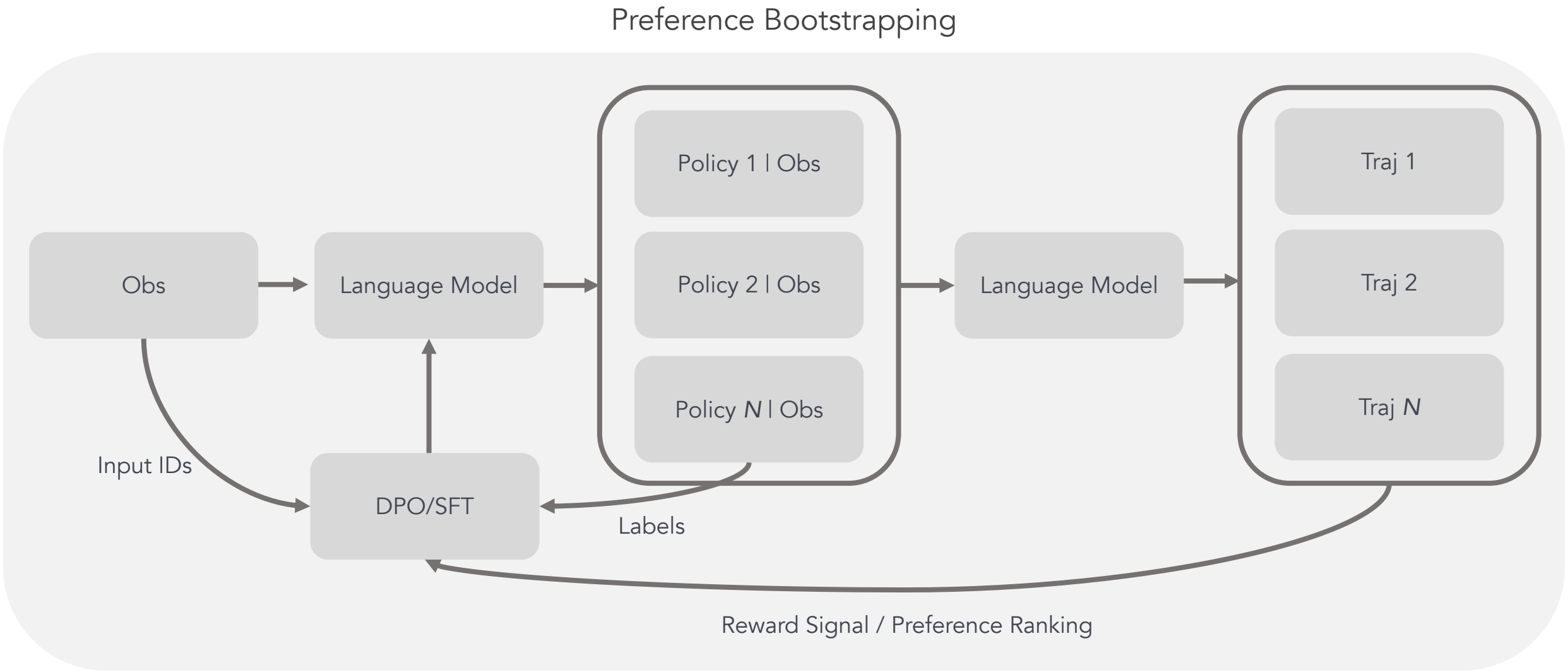
Social Contract AI (SCAI)



Social Contract AI (SCAI)



What's next?



Inspired by STaR: Bootstrapping Reasoning with Reasoning [https://arxiv.org/abs/2203.14465]

Thanks!

Jan-Philipp Fränken, Sam Kwok, Patrick Ye, Kanishk Gandhi
Dilip Arumugam, Jared Moore, Alex Tamkin

Tobias Gerstenberg, Noah Goodman