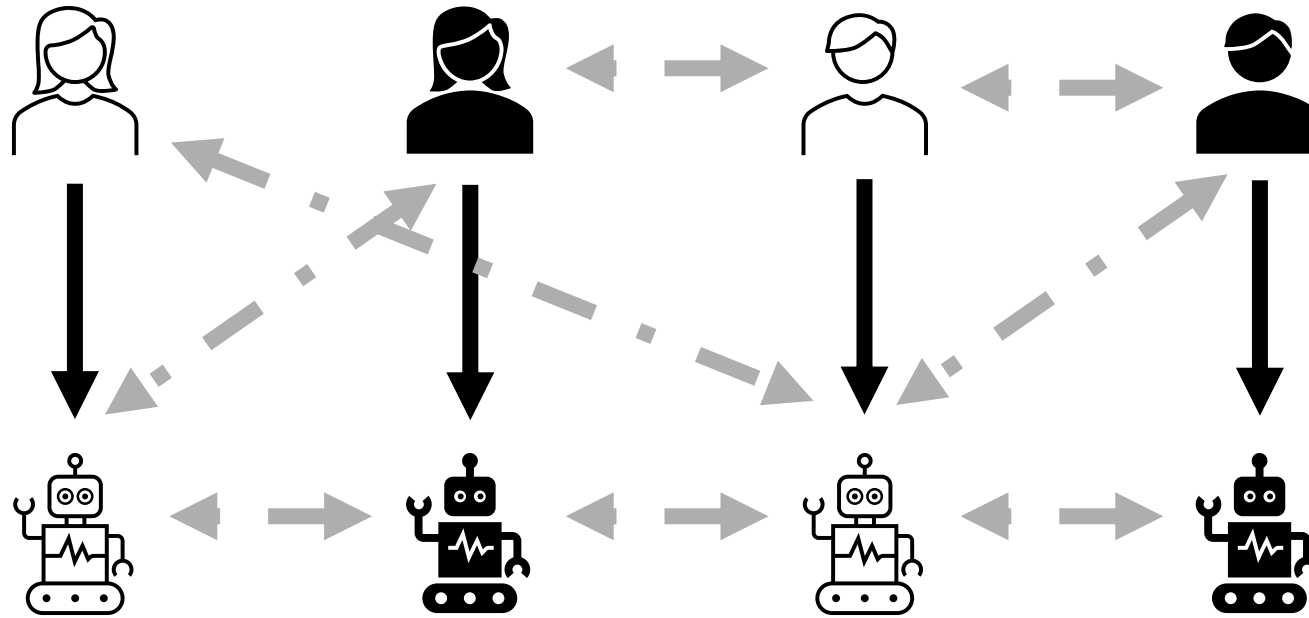


Agentic AI Design Should be Mediated to Promote Social Welfare

Omer Madmon and Moshe Tennenholtz

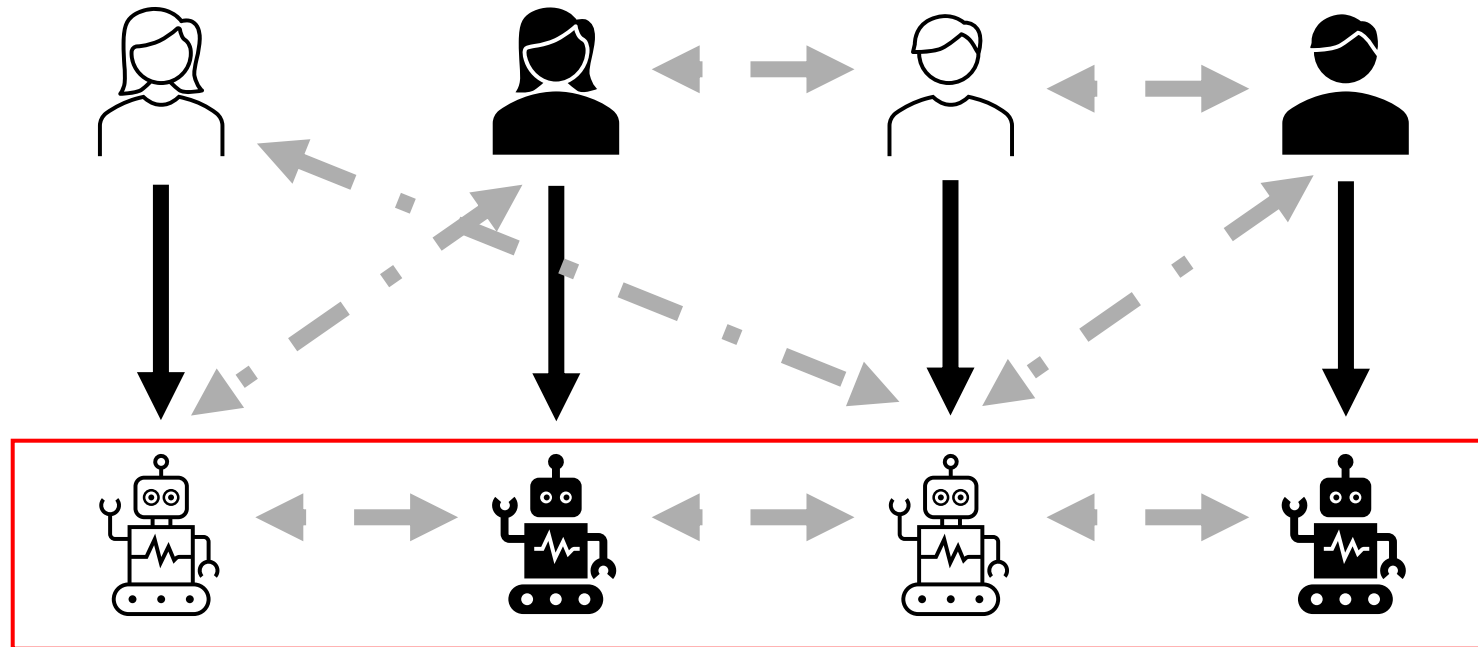


The agentic AI ecosystem



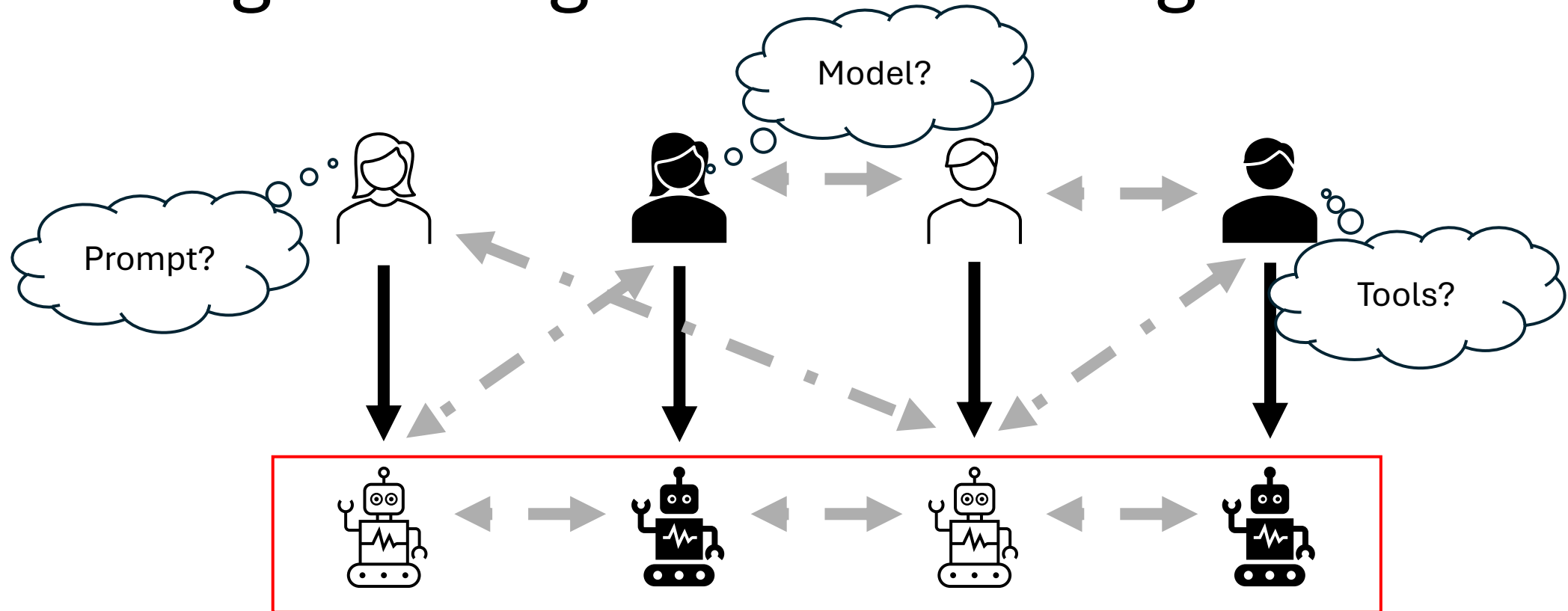
How do we make sure the interaction outcome aligns with societal objectives?

Mediators for AI agents



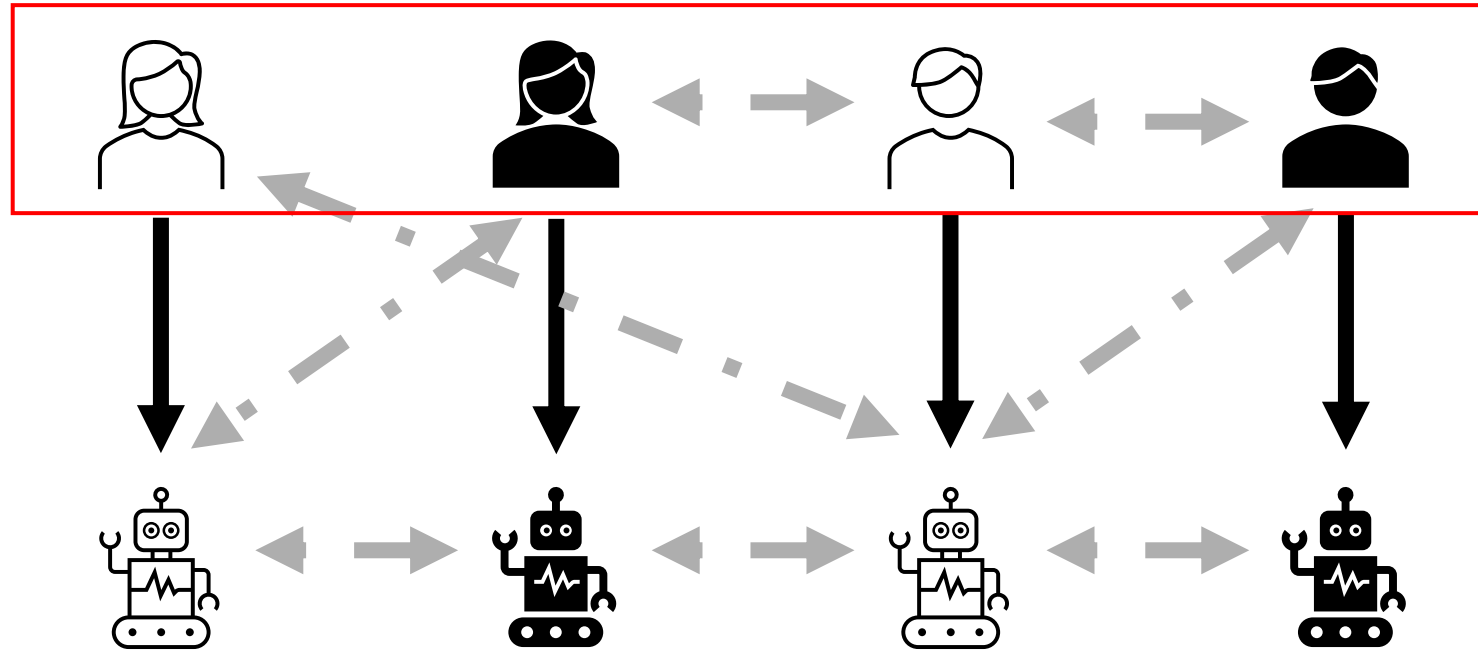
Most of the literature: apply game-theoretic mediators to the (downstream) agents' interaction

But agent design is also a strategic choice



Most of the literature: apply game-theoretic mediators to the (downstream) agents' interaction

Hence, it should be mediated as well!



Our position: we should also mediate the
(upstream) agent design meta-game

Contributions

- Discuss theoretical frameworks for agent design mediators
 - Strategy space restriction
 - Monetary payments
 - Correlation devices
 - Information design

Contributions

- Discuss theoretical frameworks for agent design mediators
- Demonstrate their effectiveness on a real-world negotiation agents design
 - Social welfare can be improved by up to $\approx 29\%$ when mediating the agent design meta-game!

Contributions

- Discuss theoretical frameworks for agent design mediators
- Demonstrate their effectiveness on a real-world negotiation agents design
- Propose a concrete research agenda

Contributions

- Discuss theoretical frameworks for agent design mediators
- Demonstrate their effectiveness on a real-world negotiation agents design
- Propose a concrete research agenda
- Discuss implications, limitations, and alternative views

Looking forward to discussing these ideas!