Sufficient Conditions for Cooperation Between Rational Agents

Anthony DiGiovanni and Jesse Clifton

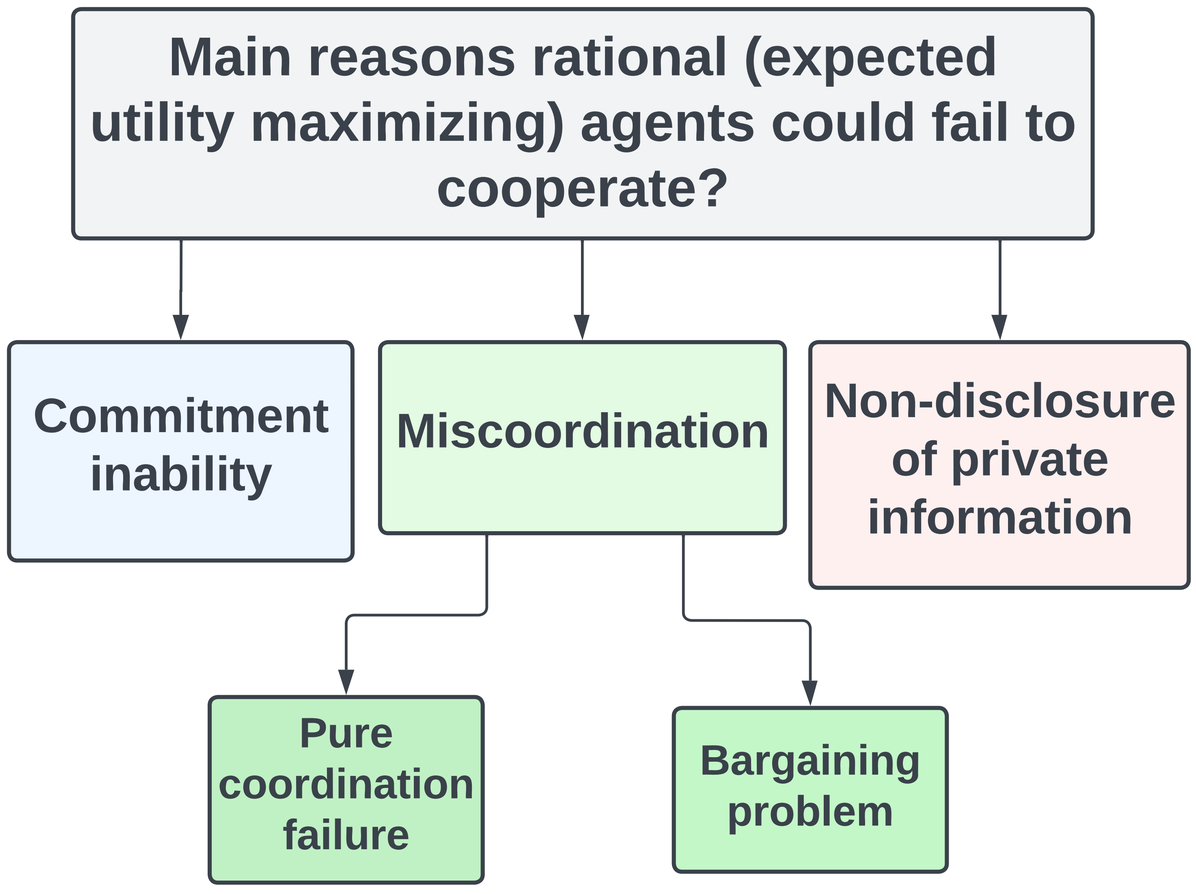
CENTER ON LONG-TERM RISK

Introduction

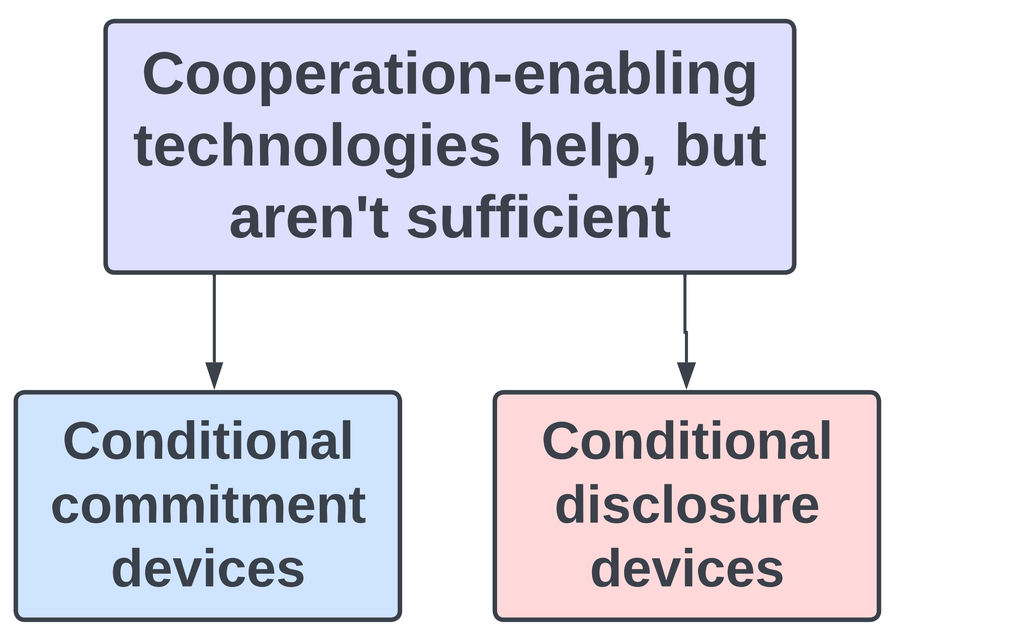
- **Cooperative AI:** design AIs that achieve socially efficient outcomes when making decisions for humans
- What are all possible **reasons even very intelligent AIs might not cooperate** (i.e., reach Pareto efficiency)?
- Some causes of cooperation failure ("conflict") are well-studied
  - AIs could overcome these with **cooperation-enabling technologies** like conditional commitment devices
  - *On-equilibrium* causes: not exhaustive!
- **Goals:**
  1. Taxonomy of all causes of conflict including rational off-equilibrium play
  2. Framework for on-equilibrium causes identifying which cooperation-enabling technologies can solve them

Framework and Examples

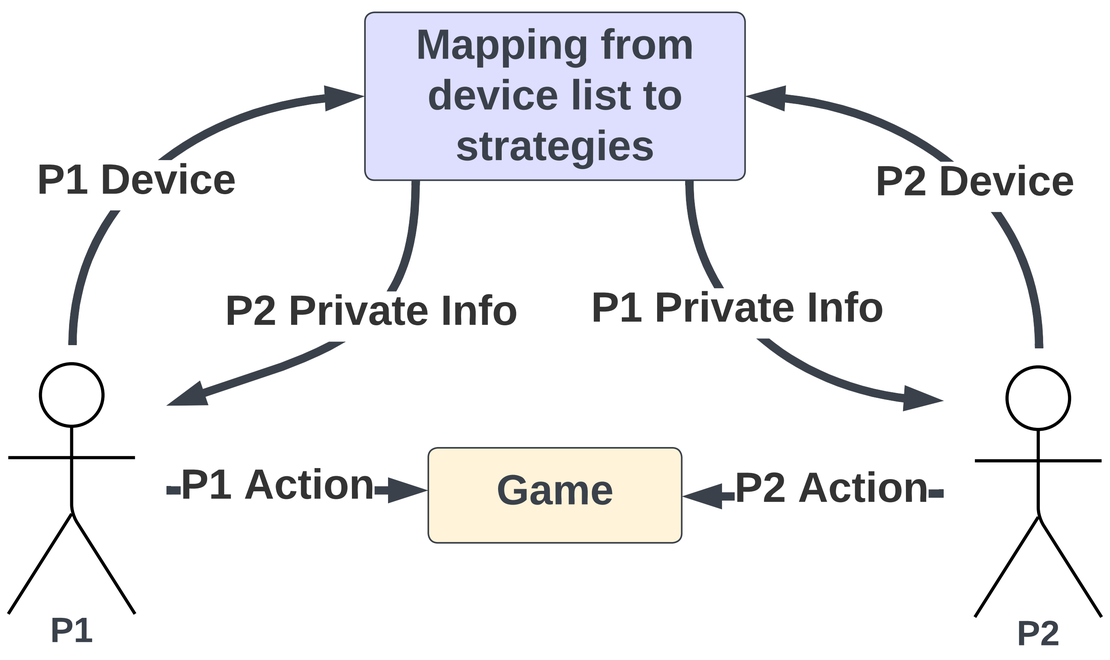
- **Credible commitment inability:** All Nash equilibria inefficient + my cooperation can't be made conditional on yours
  - *Ex:* Prisoner's Dilemma
- **Non-disclosure of private information:** My uncertainty about you makes cooperation irrational + you can't/won't resolve that uncertainty
  - *Ex:* Seller hides their valuation of a product
- **Miscoordination:** We both try to maximize expected utility by playing the same equilibrium, but our beliefs lead to playing strategies from different equilibria
  - **Pure coordination failure:** We both prefer the same outcome
    - *Ex:* Schelling NYC game
  - **Bargaining problem:** The best possible outcome for me isn't the best for you
    - *Ex:* Chicken



First exhaustive taxonomy of causes of rational conflict



Solving coordination problems is a key priority for Cooperative AI



Take a picture to download the full paper

More Details

Conditional commitment devices

- (Including commitment to following randomization)
- My device commits me to cooperate with you *if and only if* your device is one that would cooperate with me
- $\Rightarrow$ Efficient equilibrium always exists **without private information** [Kalai et al., 2010]
- **Implementation?**
  - Robust program equilibrium: Programs recursively call each other + random cooperation [Oesterheld, 2019]

Conditional disclosure devices

- My device commits me to share my private info *if and only if*:
  1. Your disclosure device is one that would share your private info
  2. Your commitment device is one that would cooperate with me
- $\Rightarrow$ Efficient equilibrium always exists **even with private information** [DiGiovanni and Clifton, 2022]
- **Implementation?**
  - Modular AI architecture, "utility function" separate from module implementing the commitment
  - Secure simulator where AIs verify each other's code + can't leak unauthorized info

Future Directions

- Causes of **more or less severe** inefficiency
  - Safe Pareto improvements [Oesterheld and Conitzer, 2021]: prevents particularly bad inefficiencies
- **Interactions between different causes** of inefficiency