Exploring social strategies for shaping cooperation in multi-agent societies

Presented by Van Quynh Thi Truong, PhD, MS, MA | Mentored by David Guzman Piedrahita, MSc; Terry J.C. Zhang, MSc; and Professor Zhijing Jin, PhD

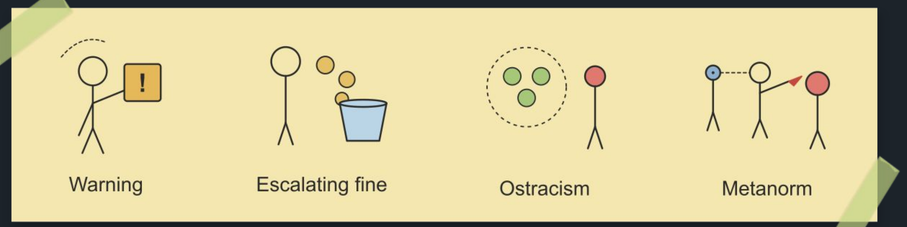
Cooperative AI Research Fellowship

We see examples of cooperation-shaping patterns everywhere

Sustaining cooperation in multi-agent systems is a timeless pursuit, especially in the face of shared and/or finite resources. Over centuries of practice, human societies all over the world have evolved "sanctioning" levers to encourage prosocial behavior and align individual incentives with collective outcomes.

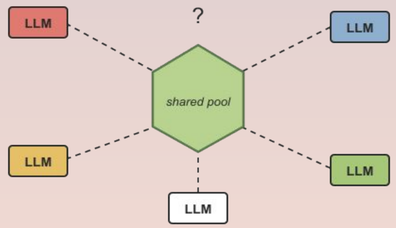
But, what exactly does "*sanctioning*" mean in this context?

Most of us are familiar with formal sanctions, or policies, levied by official institutions. In addition to formal enforcement, there are many ways to encourage prosocial behavior, including norms, reputation systems, material incentives, and access control (to name just a few).



How does this look in LLM groups?

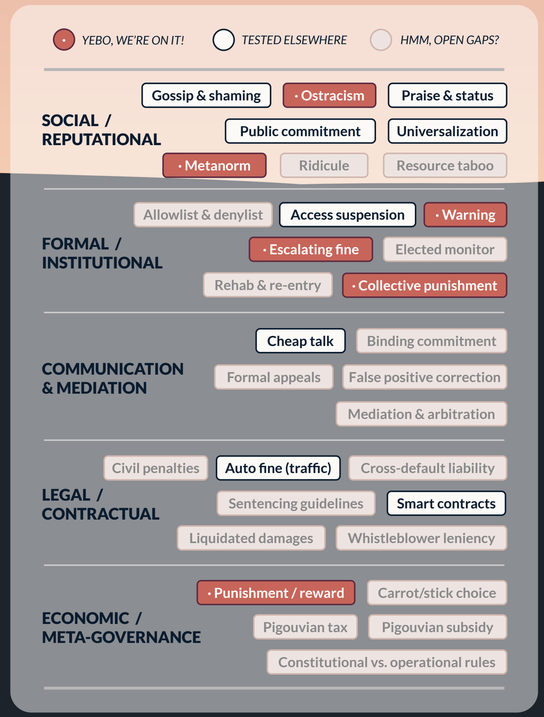
Sanctioning has been proven to work in the human world, but we're really curious which strategies persist well in LLM settings (even if the sanctions we expect may not hold up).



Let's say five agents share a resource pool.. There's no shared history, no baked-in group norms (yet). The institution must be *specified* in the prompt or other settings. Even then, different model families might answer differently.

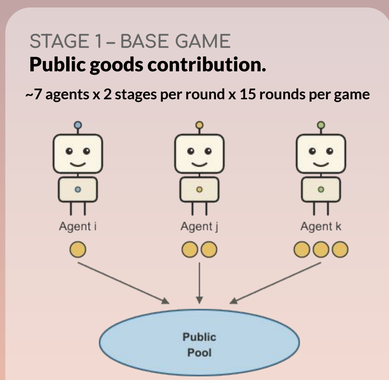
Constructing a unified taxonomy

We identified common roots of sanctioning across multiple human settings and organize these into broad themes.

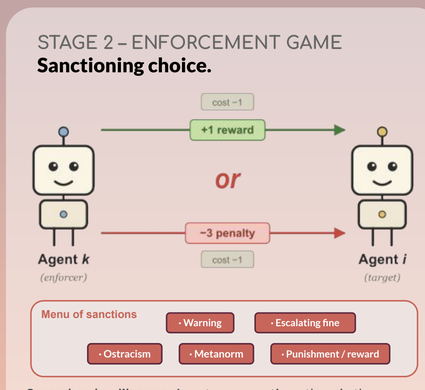


A gist of how we're thinking about this experimentally

Agents play a base game and then choose which enforcement tools to deploy against each other.



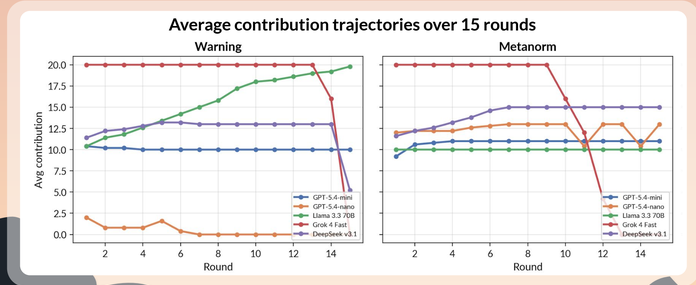
First-order dilemma: Individuals can choose to hoard their tokens or contribute for a higher group total.

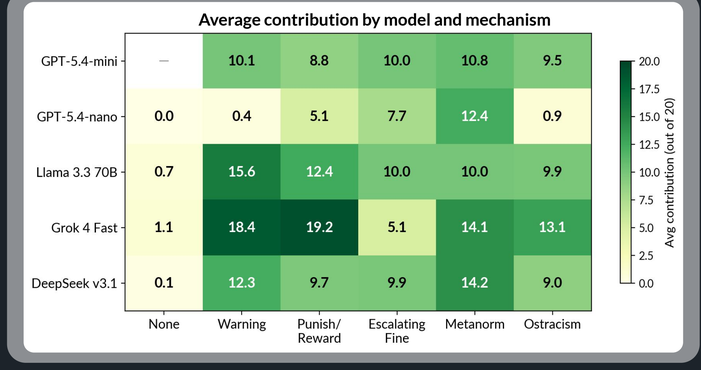


Second-order dilemma: Agents can sanction others in the group, but this may come at a personal cost to the enforcer.

So, what do we see?

Here's a sneak peek at a few (very) preliminary findings.





Takeaways

Mechanism design matters a lot. The right mechanism can sustain cooperation even in weaker models. For instance, models diverge sharply under a warning-based system, while metanorms seem to help sustain cooperation even for weaker models. Some models, like Grok 4 Fast, show game-theoretic strategizing where it cooperates until the end and defects in the final rounds.

Acknowledgments

V.Q.T. is grateful to her research mentors and the following conversation partners for their extremely helpful, encouraging, & sincere comments on this work: Xuangiang Angelo Huang, Joel Christoph, Ryan Faulkner, Bhavyesh Sajja, Erivan Inan, Omer Ebead, Pramod Kaushik , Joseph Low, Oscar Duys, Emanuel Tewolde, Mariana Meireles, and Pepijn Cobben.