Multi-Agent AI Systems Need Institutional Design, *Not Just Model-Level Alignment*

*Equal co-first · †Equal co-last


*Van QT Truong, PhD


X. Angelo Huang


Erivan Inan


Ryan Faulkner


Joel N. Christoph


†Terry JC Zhang


†David Guzman Piedrahita


†Zhijing Jin, PhD

Spotlight Paper!

ICML International Conference On Machine Learning

AI4GOOD Workshop

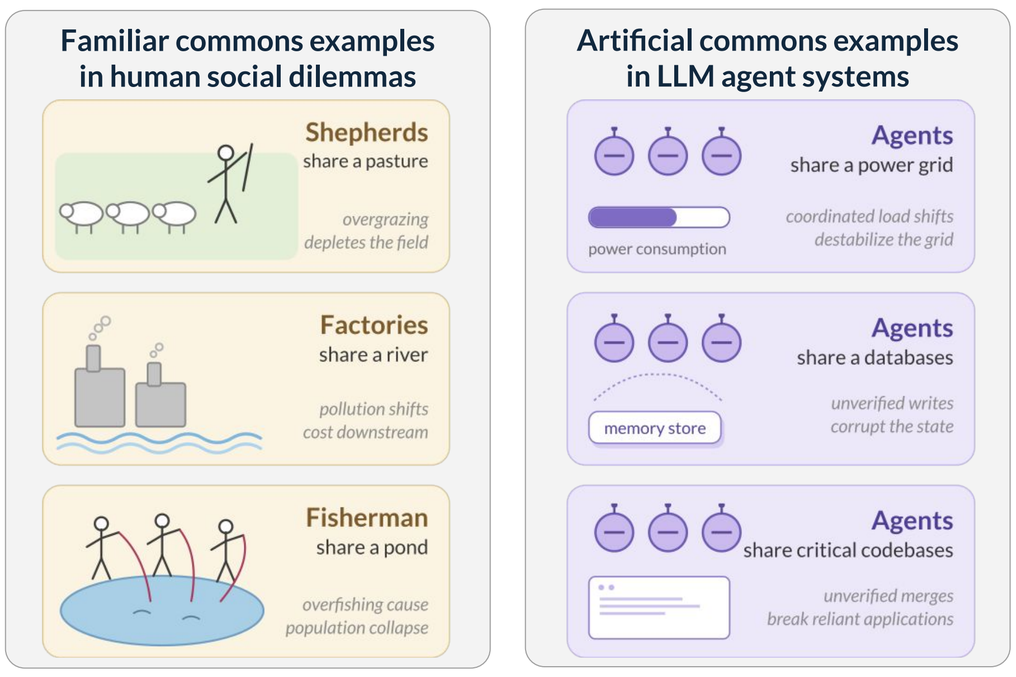
OUR POSITION

Multi-agent AI safety requires intentional institutional design. Even individually aligned agents can produce *collectively unsafe* outcomes when sharing finite resources. We believe agent systems should be studied as **governed systems** whose institutions shape collective behavior.

PROBLEM

LLM agents are already embedded in real-world settings.

Everyday, teams of sub-agents, on your computer or on the Internet, must vy for and share finite tools and resources. These shared resources, like open-source codebases, public databases, or compute (among many others!), form *artificial* commons, where classic collective-action failures reappear.

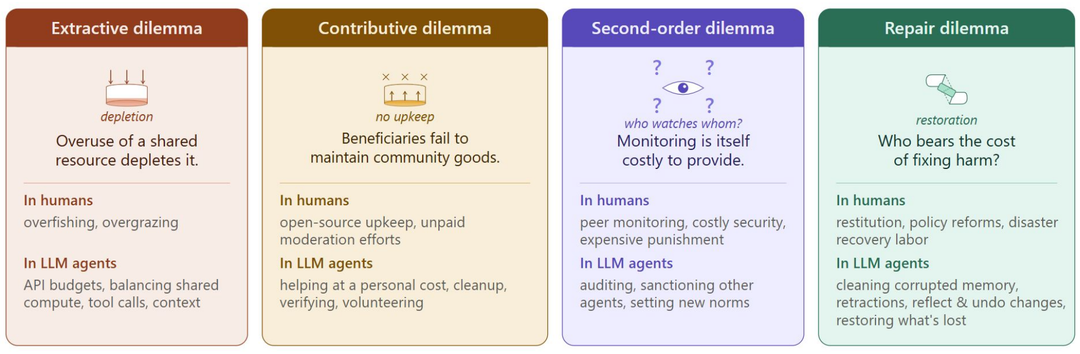


But, aligned individual agents do not simply add up to a cooperative society.

OUR CURRENT FRAMEWORK

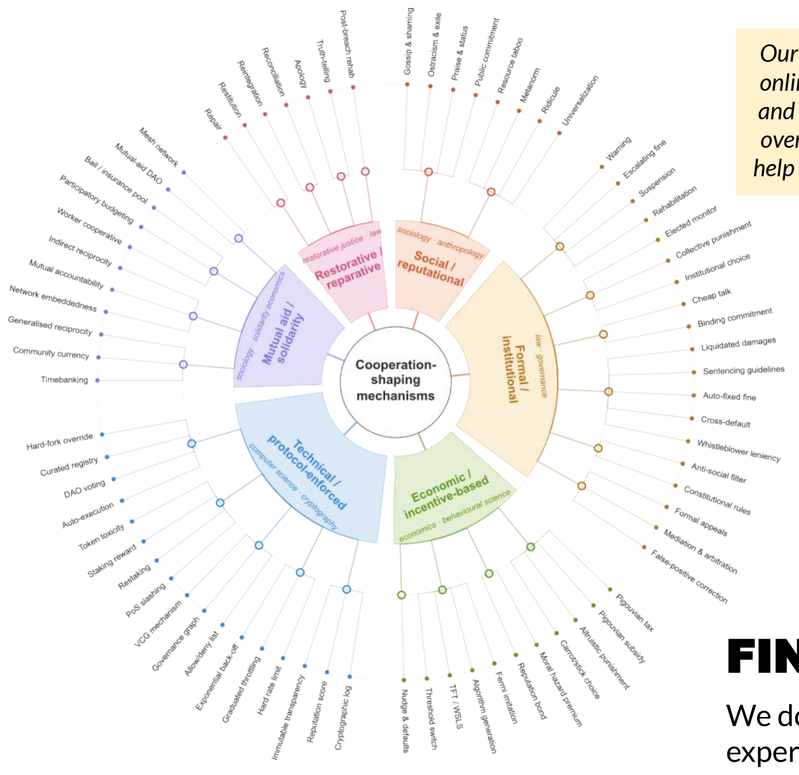
Aligned individuals are not enough; group-level design is needed.

We start by outlining 4 types of cooperation dilemmas in human and LLM-agent systems. There are likely many, many more that are worth discussing too!



We mapped what's in the literature to a unified taxonomy.

When we looked at the vast literature, we noticed similar patterns across disciplines: legal/justice settings, economics, IT/networking, anthropology, community practices, and more. Finding it useful to see altogether, we mapped countless examples of sanctioning, policy, social norms, mechanism design, and interaction protocols into a single diagram. Below is our *work-in-progress* visual.



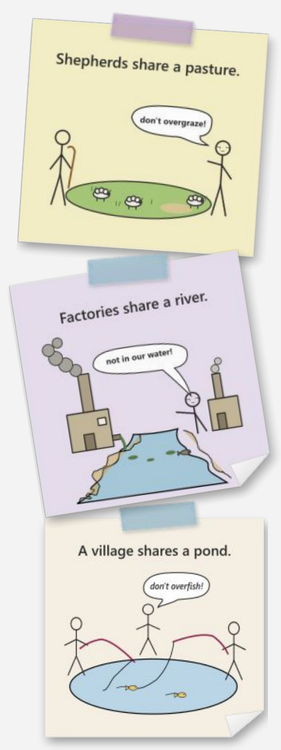
Our plan is to publish this online with clickable links and welcome feedback on overall structure, form, or help to spot missing pieces.

MOTIVATIONAL DESIGN SPACE

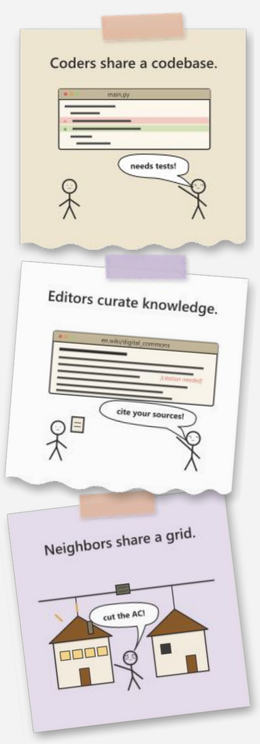
We see examples of cooperation-shaping patterns everywhere.

Sustaining cooperation in multi-agent systems is a timeless pursuit, especially those living through times of scarcity. Over centuries of practice, human societies all over the world have designed mechanisms, sanctions, or other levers to encourage prosocial behavior and align individual incentives with collective outcomes. Some for the greater good, and others as coercive regimes.

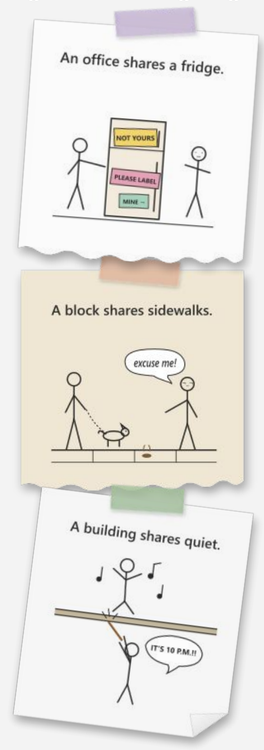
These *tragedy of the commons* scenarios are well-studied in economics & game theory.



And it's fair to wonder how this plays out across our digital & physical worlds?



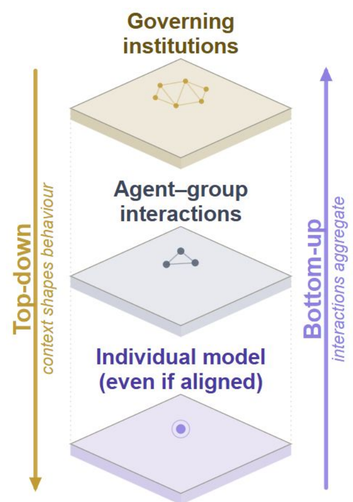
Especially given that us humans are not perfect at living or working together.

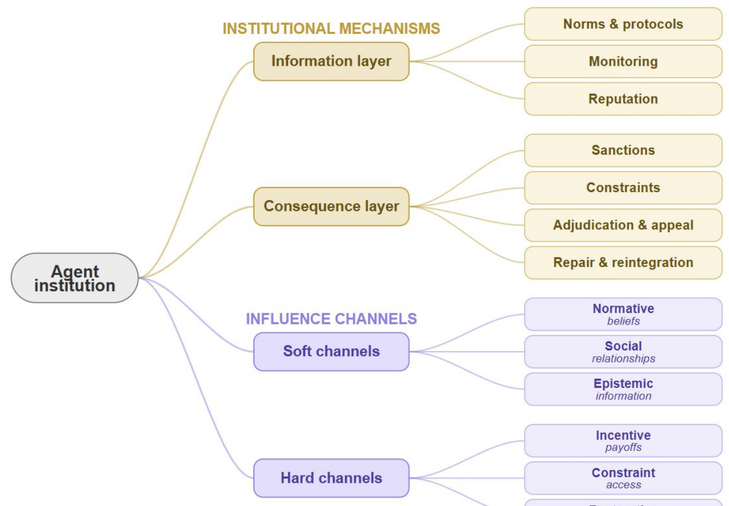


CHARTING A PROMISING RESEARCH AGENDA

We looked at possible design dimensions for AI agent groups and institutions.

We wanted to tease out ways researchers can observe and influence emergent collective dynamics, including thinking about what an institution must do and mechanism families for such processes.





It is worth asking: what are the best ways to capture cooperative outcomes?

Capturing cooperative outcomes may require empirical, quantitative, qualitative, and interpretive methods. The practical question is which interaction levels and metrics are most feasible to evaluate, since group- and population-level interactions quickly become combinatorially complex and costly.

1 Observer	2 Arbiter	3 Enforcer
Who can see relevant behavior?	*Who decides whether a violation occurred?*	*Who applies the sanction?*
In LLM agents: peer agents, logs, monitoring systems, human auditors	**In LLM agents**: rule-based checker, mediator agent, judge, voting system, human review	**In LLM agents**: peer agents, platform, tool orchestration layers, authority figures, smart contracts

Timing	Reversibility	Adaptivity	Observability	Gameability	Repair	Scale
before or after harm?	*can it be undone?*	*does it escalate?*	*visible or hidden?*	*can it be exploited?*	*does it restore?*	*across many agents?*
ex ante rate limits vs. ex post fines	warning vs. permanent exclusion	graduated sanctions, re-entry conditions	tool logs vs. private reasoning	reputation farming, surface compliance	memory cleanup, output correction	local trust vs. platform-wide

FINAL PARTING WORDS

We do not yet have all the answers for this research agenda, but we are excited by its promise and look forward to sharing ongoing experiments and collaborating with others!

So long as humans bear the consequences of AI behavior, we believe humans should have a say in which values are embedded in AI agent institutions, even if capable AI systems can eventually design and run those institutions themselves. Those choices may be amplified in future AI agent economies and societies, including fully automated systems that no longer require humans in the loop.

Please reach out if you ever want to chat!

Van QT: scientistvan@gmail.com
Angelo: hxuanqiang@ethz.ch
Zhijing: zjin@cs.toronto.edu

UNIVERSITY OF TORONTO · VECTOR INSTITUTE | INSTITUT VECTEUR · EuroSafeAI · ETH zürich · IDAI · Max Planck Institute for Intelligent Systems