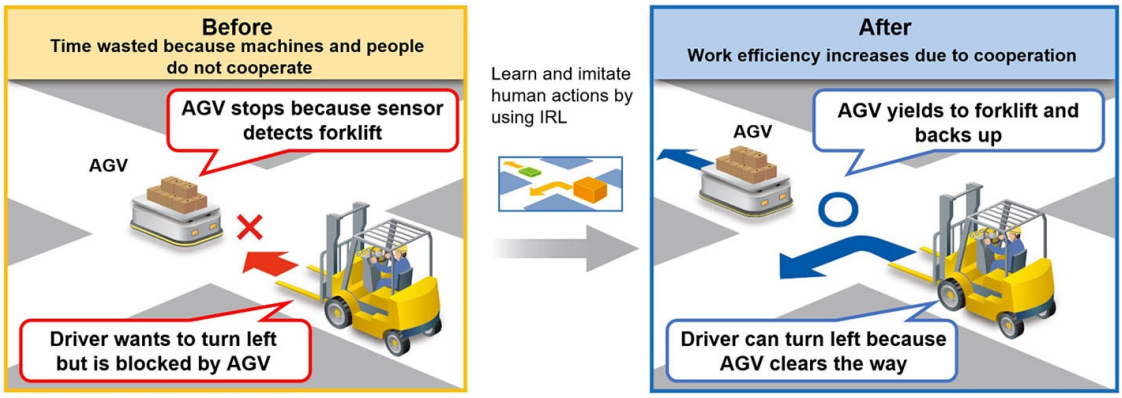
Problem Set 2: Cooperative AI and the Future of Mechanism Design

> *Author: Habib, Debaya*

> *Class, Major, and Affiliation: Class of 2025, Mathematics, Duke Kunshan University*

> *Disclaimer: Submissions to Problem Set 2 for COMPSCI/ECON 206 Computational Microeconomics, 2022 Spring Term (Seven Week - Second) instructed by Prof. Luyao Zhang at Duke Kunshan University.*

RQ1: What is cooperative AI? What do you see as the potential that computer science and economics can jointly contribute to advance cooperative AI?



(Mitsubishi 2020)

According to Cooperative AI's website, Cooperative AI aims to reconceive AI as **deeply social** (Dafoe 2021). More practically, cooperative AI wants to find ways where AI is not engaged in a zero-sum conflict game but rather a game where common and conflicting interests are present and AI has to find ways to work with the opposite party, whether it is another AI or a human. I think the contribution of Computer Science will center on the AI algorithms. Better deep learning models will allow AI to process more information at higher speeds so that it can make more accurate information faster. As for Economics, I believe that the game theory models will be very useful in guiding AI towards the best cooperative decisions to make during games.

RQ2: Besides the desirable outcomes of cooperation, what other desirable outcomes the mechanism design theory aims at achieving?

Mechanism design, at its heart, wants to allow agents who are self-interested and with incomplete information to make decisions that create utilitarian good (Chen 2022). I believe that Mechanism Design (MD) is a very virtuous field of study as it finds ways to channel selfish human behavior to achieve positive outcomes. MD regulates specific mechanisms related in the game as to engineer specific outcomes. In this sense, MD is often referred to as *reverse game theory*.

RQ3: What are the limitations of the current mechanism design theory in achieving desirable outcomes? What new challenging are we facing in a new era with more and more human and AI interactions?

Impossibility theorems, namely Arrow's impossibility theorem and the Gibbard–Satterthwaite theorem, represent limitations to current mechanism design theory. Both theorems revolve around the fact that voting systems may not produce the desired result of voters even if every voter votes sincerely.[^1] [^2] These impossibility theorems present an issue to Mechanism Design. Since MD attempts to devise systems that produce positive results in spite of self-interest and incomplete information, impossibility theories imply that Mechanism Design cannot always achieve its stated goals. With more human and AI interactions, we might find ourselves in new games that are certain to produce results that are not ideal. As such, we should not only advance the development of cooperative AI but also make sure that the games we play have a possibility of achievieng desirable outcomes.

[^1]: Liberto, “What Is Arrow's Impossibility Theorem?,” 2021. [^2]: Svensson, “The Proof of the Gibbard–Satterthwaite Theorem Revisited,” 2014.

RQ4: What new challenging are we facing in a new era with more and more human and AI interactions?

Dafoe et al. (2020) defines the main challenges that we face:

1. Improvement in cooperative abilities can be an exclusionary force and/or encourage cooperative agents to work together to the detriment of other agents.
2. The better agents get at cooperating, the more likely they are to possess a skillset that allows them to better deceive.
3. Cooperative is fundamentally linked to coercion and competition

Fundamentally, the issues with cooperation with AI and humans is that not all forms of cooperation are created equal. We must work diligently to make sure that we are cooperating towards positive means. Perhaps even semantically, the word cooperation has an inherent positive connation. However, it is wise to watch out for uses of AI to help humans achieve undesirable outcomes: taking value from society, cheating, etc.

Additionally, coercion and cooperation share the same skillset giving agents who engage in socially encouraged cooperative activities the ability to cause outsized harm to society, embodied in their propensity for coercion. In the same vein as making sure that we are doing good cooperation, we should also ensure that cooperative agents know that coercion is a punishable offence and that their skillset should only be channeled towards positive causes.

Nonetheless, Dafoe et al. remain optimistic on our ability to find solutions to these challenges. They conclude their research paper with the following:

> *As the field of AI takes increasingly confident strides in its ambition to build intelligent machine agents, it is critical to attend to the kinds of intelligence humanity most needs. Necessarily among these is cooperative intelligence.*

RQ5: Based on your interest and your advantage in skills, how do you plan to contribute to overcoming the limitations of the current mechanism design theory in achieving the desired outcomes and making this world a better place?

I believe I can help in two ways.

- First, and especially as I gain more technical skills, I think I can help develop mechanism design models. I personally believe that the field of game theory in general has incredible upside in streamlining human society towards better achieving progress. There are a lot of smart and ambitious people in the

world who are not given the chance to excel because the institutions around them are barring them from opportunities. This can be because the incumbents do not want to be unseated by competition. And sometimes it is due to simple incompetence. Creating systems that produce positive outcomes will hone the machinery of society.

- Second, I believe that I can find ways to advocate for the Mechanism Design theory's merits to people around me to get more interest in the field. I believe that the more that society is aware of the need to design systems that produce positive outcomes (as opposed to just going at it and hoping for the best), the more likely we are to see progress occur.

Finally, it is worthwhile to touch upon the prospects of advancing game theory and mechanism design as the two fields seem to be stuck as of late.

I believe the best way to advance both fields is to look at quantum computing. It has been proven that quantum algorithms perform much better in classical strategies such as coin-flip games and the prisoner's dilemma (Allen 2020). Therefore, quantum computing, through divising better algorithms, presents an opportunity to advance the field in new ways. As for mechanism design, Quantum mechanism design is a natural progression of quantum game theory. Since mechanism design is essentially reverse game theory, quantum theroy must be in a position to provide advances to mechanism design. An example is that an agent that completes a condition can tackle "bad" social choice rules evading the restrictions of traditional mechanism design theory.

Glossary

Term	Definition	Source
Mechanism Design	Mechanism design theory is an economic theory that seeks to study the mechanisms by which a particular outcome or result can be achieved.	Chen 2022
Arrow's Impossibility Theorem	Arrow's impossibility theorem is a social-choice paradox illustrating the flaws of ranked voting systems. It states that a clear order of preferences cannot be determined while adhering to mandatory principles of fair voting procedures.	Liberto 2021
Gibbard–Satterthwaite Theorem	The Gibbard–Satterthwaite theorem states that with three or more eligible alternatives, a voting rule is strategy-proof only if it is dictatorial.	Svensson 2014

References

Allen, Khaled. "An Exploration of Quantum Game Theory and its Applications." 2020. University of Colorado

Corporation, Mitsubishi Electric. 2022. “Mitsubishi Electric Develops Cooperative AI for Human-Machine Work: 2020: Global News.” MITSUBISHI ELECTRIC BELGIUM. Accessed May 3. https://be.mitsubishielectric.com/en/news/releases/global/2020/0603-a/index.html.

Dafoe, Allan, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. "Open problems in cooperative ai." arXiv preprint arXiv:2012.08630. https://arxiv.org/abs/2012.08630

Dafoe, Allan, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. “Cooperative AI: Machines Must Learn to Find Common Ground.” Nature News. Nature Publishing Group. May 4. https://www.nature.com/articles/d41586-021-01170-0.

Chen, James. 2022. “Mechanism Design Theory.” Investopedia. Investopedia. February 8. https://www.investopedia.com/terms/m/mechanism-design-theory.asp.

Liberto, Daniel. 2021. “What Is Arrow's Impossibility Theorem?” Investopedia. Investopedia. July 21. https://www.investopedia.com/terms/a/arrows-impossibility-theorem.asp.

Svensson, Lars-Gunnar, and Alexander Reffgen. 2014. “The Proof of the Gibbard–Satterthwaite Theorem Revisited.” Journal of Mathematical Economics 55: 11–14. doi:10.1016/j.jmateco.2014.09.007. https://www.sciencedirect.com/science/article/abs/pii/S0304406814001177