Dissertation proposal: New foundational ideas for cooperative AI

Caspar Oesterheld

February 2025

Abstract

My doctoral research addresses two fundamental obstacles to beneficial outcomes from strategic interactions between multiple parties: strategic incentives against cooperation (as in the Prisoner's Dilemma) and the multiplicity of solutions (sometimes called the equilibrium selection problem). As AI systems are increasingly involved in consequential decision making processes on behalf of human principals, understanding how to achieve desirable outcomes in multi-agent AI settings becomes critical. My research leverages unique features of AI systems – including their transparency, reproducibility, and malleability – to develop novel game-theoretic approaches that enable better, more cooperative outcomes.

Three primary research directions form the core of this dissertation. First, the concept of "safe Pareto improvements" provides a rigorous framework for improving outcomes without resolving equilibrium selection problems. Unlike traditional solution concepts, safe Pareto improvements make qualitative assumptions about pairs of games rather than individual games. This sometimes allows us to prefer playing one game over another, without any judgment about how each of the individual games is played. Second, the concept of *program equilibrium* explores how the use of mutually transparent decision-making algorithms can allow for cooperation. Third, my research on so-called *Newcomb-like* decision problems takes inspiration from philosophical branches of decision theory. I investigate how cooperation can be achieved when different parties deploy similar AI systems.

Current and planned work extends these directions through several projects, including: connecting program equilibrium with mediated equilibrium; exploring sequential program/mediated equilibrium-type settings; investigating the relationship between self-locating beliefs and decision theory; developing theoretical foundations for safe Pareto improvements, as well as analyzing safe Pareto improvements in a new setting. I've also started to implement some of these theoretical ideas in language models to test their practical applicability.

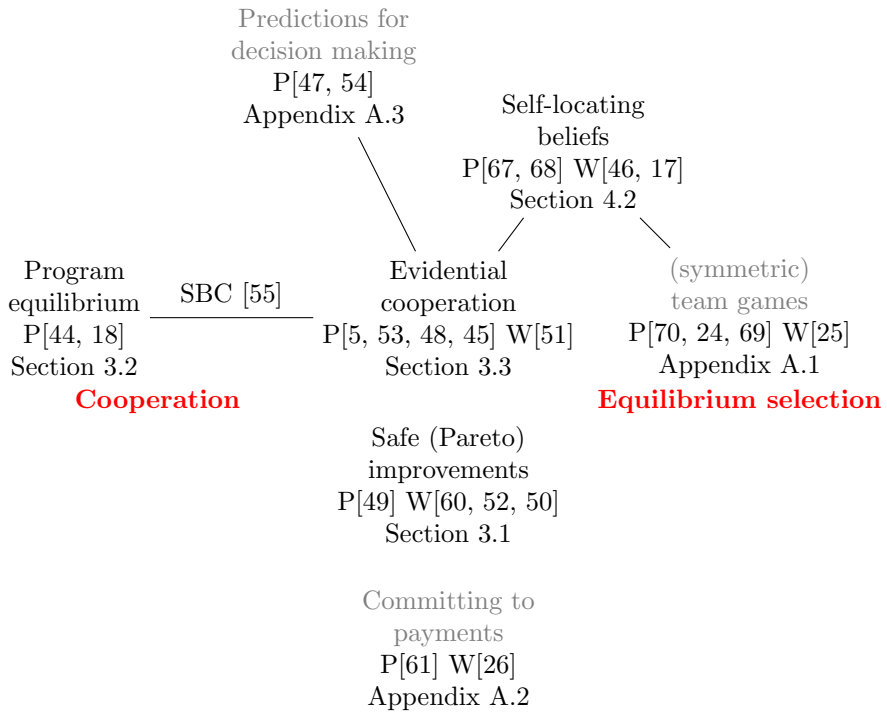

Figure 1: The figure gives an overview of the two central problems and the main approaches to these problems that I have pursued in my doctoral work. Connections between the areas are highlighted by edges between them.

1 Introduction

In this document, I outline the main directions of my doctoral research. I start in Section 2 by describing the problems that my research aims to address. I then continue to briefly describe the three most central directions themselves in Section 3, generally with a focus on completed work. Most of my dissertation will center on these directions. Finally, in Section 4, I described current and future work that I plan to feature centrally in the dissertation.

In Appendix A, I discuss some further directions and lines of work and how they connect to the main problems and directions – I expect these to take up less space.

Last and least, since the CMU CSD PhD student handbook instructs me to "[d]emonstrate [my] personal qualifications for doing the proposed work" [2, p. 35], I include some biographical information in support in Appendix B.

Related work This document outlines a research agenda toward improving outcomes from interactions between multiple AI systems. Other such agendas

have been put forward. Most closely related is one I have co-written:

> Vincent Conitzer and Caspar Oesterheld. Foundations of cooperative AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15359–15367, Sep. 2023. doi: 10.1609/aaai.v37i13.26791. URL `https://ojs.aaai.org/index.php/AAAI/article/view/26791`

In particular, the introduction and the section "Equilibrium Selection" describe essentially the same problems as Section 2. The sections "Cooperation by Reading Each Other's Code" and "Cooperation between Copies" describe the same approaches as Section 3.2 and Section 3.3, respectively. I also address the issues discussed in "Self-Locating Beliefs" in some of my current/future work, see Section 4. Others have promoted agendas toward similar problems (improving outcomes of multi-agent issues in general) with varying levels of overlap with the present agenda [20, 13, 22, 28]. Finally, a few authors have advocated some similar research directions with somewhat different AI-related motivations. For instance, Soares and Fallenstein [64] have advocated for investigating the research direction I discuss in Section 3.3. (Cf. Conitzer [15].)

2 The problems I'm aiming to address

2.1 Two (known) obstacles to cooperation

Strategic incentives against cooperation When two or more agents interact, an individual's preferences over her possible courses of actions may often be irreconcilably misaligned with other agents' or social preferences (i.e., with preferences aggregated across the agents). The textbook example for this phenomenon is the famous Prisoner's Dilemma, wherein each individual participant prefers Defect over Cooperate, but social preferences favor mutual cooperation over mutual defection. Essentially the same phenomenon is also known as the Tragedy of the Commons. Besides these most basic examples, various more complicated dynamics can arise, as illustrated by some of the classic examples in game theory, like the Traveler's Dilemma, the Trust Game, the Security Dilemma.

In all of these interactions, some form of cooperation is desirable (to the participants in the interaction), but not achieved by default if the players act rationally. Like many others, I'm interested in how to add further twists to these stories in order to allow for cooperation.

Equilibrium selection A strategic interaction can have multiple solutions, e.g., Nash equilibria. A paradigmatic example of equilibrium selection is the Game of Chicken. A more realistic example is the negotiation of any beneficial deal between a group of parties.

The multiplicity of strategic solutions poses two problems: First, some equilibria might be more desirable than others. (The Stag Hunt is the canonical example of such a game.) Second, we might worry that agents might play the

actions from different Nash equilibria. For example, in the Game of Chicken, we might worry that both players will go Straight for their favorite equilibrium and thus crash, obtaining the worst possible utility for both players.

Compared to the problem of the Prisoner's Dilemma, the *generic* equilibrium selection problem (given an arbitrary normal-form game, what equilibrium (if any) should be played?) has received much less attention.

2.2 The AI angle

In my PhD I've specifically been motivated by multi-agent interactions involving AI, such as interactions between AI agents acting on behalf of human principals. As technology advances, multi-agent interactions will become increasingly common and consequential. All the same failure modes afflicting interactions between humans and human organizations apply just as much to interactions involving AI systems. Yet, they have received relatively little study. Of course, many of the traditional ideas in game theory apply just as well to interactions involving AI as they apply to their original (or traditional) subject. For instance, tit-for-tat-style cooperation in repeated games works just as well (or better) when AI or software agents are involved as they are in interactions between unassisted humans. In my work, I therefore focus specifically on features of AI that could cause interactions involving AI systems to systematically differ from humans.

The most important for my work are the following:

- The AI design perspective forces us to give definite, complete, practical answers. For example, the typical game theory textbook will mostly not try to address the equilibrium selection problem. If we build a software agent, we are forced to decide somehow (e.g., by choice of some learning algorithm) how the agent is to resolve the equilibrium selection problem. Similarly, it may be legitimate in many contexts to simply assume it as a given that an agent has some kinds of beliefs.

- Software reasoning is potentially transparent and reproducible. If a decision is made by a particular piece of computer code, then others can read it, test its behavior for various inputs and perhaps understand how it works. In contrast, as a human I usually cannot fully reveal how I'm going to make decisions (unless I myself do little more than implement a simple algorithm).

- It may be common for AI decision making algorithms to interact with copies of themselves. For instance, right now GPT-4 makes (generally small, unimportant) decisions in a world in which other (generally small, unimportant) decisions are also made by GPT-4 (prompted to serve a different user's objectives).

- Software decision makers are much more malleable than human agents. It is widely recognized that it can sometimes be strategically beneficial to

	Dare	Chicken
Dare	$-8+b, -8+b$	$4+b, 1+b$
Chicken	$1+b, 4+b$	$2+b, 2+b$

Table 1: A Game of Chicken with a parameter b

be a mad man. For example, it might be easier to motivate an irreplaceable employees (say, the only employee who understands some essential COBOL code) if the manager is (believed to be) willing to risk the company's survival by firing the employee. I can't become a mad man at will and so the irreplaceable employee will know that I wouldn't ever let them go. But I can easily make, say, a GPT-4-based decision maker that acts madly.

Note that there are lots of other unique features of strategic interactions involving AI that are at least equally, if not more, important. For instance, even today computers can perform many kinds of calculations much faster and much cheaper than a human. Thus, we can set up, for example, market makers that no human could implement by hand. On the other hand, computers' speed at performing arbitrary calculations also makes oversight and human verification much harder. In many cases, it is (and increasingly) will be impossible to understand why a particular decision was made, and whether it was made correctly, erroneously or maliciously. While highly important, neither the optimistic nor pessimistic side of computers' speed is a central feature of my doctoral work.

3 The main directions of my past work

3.1 Safe (Pareto) improvements

The concept of safe Pareto improvements is a new game-theoretic ideas for improving outcomes in games with a multiplicity of solutions. I introduced this idea in the paper "Safe Pareto Improvements for Delegated Game Playing", published in *Autonomous Agents and Multi-Agent Systems* (with a shorter version published and presented at AAMAS 2021) [49]. For a more detailed, general introduction to safe Pareto improvement, I refer the reader to my blog post for the CMU CSD PhD program's writing requirement.[1] If you prefer an introduction to the concept written by someone else, see Clifton's [13, Sect. 4.2] or Baumann's [4] introductions to the idea of *surrogate goals*, a special case of safe Pareto improvements.

For brevity, I here give a maximally simple (almost trivial) example to illustrate the concept. Imagine a two-stage game. In the first stage, you choose between 0 and 1. Call that choice b. In the second stage you play the Game of Chicken as per Table 1.

[1] See `https://hackmd.io/PIbBmbx_QWK52cCQsyax0w`.

Should you play $b = 0$ or $b = 1$? Ordinary game-theoretic solution concepts (e.g., the concept of subgame-perfect equilibrium) will typically permit both $b = 0$ and $b = 1$. This is because (regardless of the value of b) the Chicken game has both the (Dare, Chicken) and the (Chicken, Dare) equilibrium. Perhaps the equilibrium changes (in our disfavor) depending on b?

But intuitively it seems clear that we should choose $b = 1$. After all, the Chicken game in Table 1 is strategically equivalent between the $b = 0$ and $b = 1$ case.

The concept of safe (Pareto) improvements formalizes this type of intuition. Roughly, the core idea is that we make (qualitative) assumptions about the outcomes of *pairs* of games. (In contrast, traditional solution concepts focus on making assumptions about a single game.) From these assumptions, one can then sometimes infer that one game should be played rather than another. For instance, to argue for favoring $b = 1$, it suffices to make the assumption that isomorphic games are played isomorphically (or perhaps that our beliefs about them should be isomorphic).

Compared to reliance on classic game-theoretic solution concepts (where all we know is that any Nash/correlated equilibrium or rationalizable strategy profile or whatever can happen), safe Pareto improvements can sometimes overcome indecision (while remaining very well justified). (See above.) Compared to concepts like "best/worst Nash" (which also rationalize the choice of $b = 1$ in the above example), safe Pareto improvements are more indecisive. (For most pairs of strategic interactions, neither is a safe Pareto improvement on the other.) But when safe Pareto improvements are decisive they are much more convincing (assuming the assumptions made are convincing). If playing Γ' is a safe Pareto improvement on Γ that should make us much more comfortable about favoring Γ' over Γ. If all we know is that the worst Nash is higher in Γ' than in Γ', then arguably we don't have a particularly strong reason to favor Γ' over Γ.

The most important (sub)directions for research on safe Pareto improvements are the following (see also Section 4):

- Clarifying the conceptual foundations
- Figuring out situations in which safe Pareto improvements be useful, i.e., in what cases can we intervene on games in a way that frequently allows for safe Pareto improvements? My published paper on safe Pareto improvements gives an answer: safe Pareto improvements are useful when you can instruct an agent with a utility function [49].
- Developing algorithms and complexity results for identifying safe Pareto improvements in various settings. (Again, my published paper offers some such results.)
- Addressing various practical obstacles to applying safe Pareto improvements in complex, real-world environments.

Published paper on safe Pareto improvements

- Caspar Oesterheld and Vincent Conitzer. Safe Pareto improvements for delegated game playing. *Autonomous Agents and Multi-Agent Systems*, 36(2), 2022. Article number 46

3.2 Program equilibrium

Prior work [41, 33, 58, 65, 3, 19][2] has studied the following setting: Consider any normal-form game Γ to be played by, say, two players Alice and Bob. Now imagine that instead of playing this game directly, Alice and Bob each choose a set of instructions in the form of a computer program from a set PROG_i. The players' strategies are then chosen on the players' behalf according to the instructions by interpreting the computer program. Importantly, the program chosen by Player 1 can read Player 2's source code, and *vice versa*. I will call this setting a *program game*. See Figure 2 for a visualization. I'll refer to equilibria of the program game as *program equilibria* [66].

Program games often allow for cooperation when the underlying game does not. For instance, in the Prisoner's Dilemma, the following program forms a cooperative equilibrium with itself: "If the opponent's program is equal to this program: Cooperate. Else: Defect".

Program games have been proposed as models of various different real-world strategic settings, especially interactions between software-driven decision makers. Blockchain technology allows for a close-to-literal implementation of program game via so-called *smart contracts*. But more generally program games model any interaction between software agents whose procedure is in substantial part (credibly) revealed to other agents. Beyond software agents, I would argue that program games can also model various other real-world interactions between somewhat transparent agents. For instance, Critch et al. [21] use program games as a model of interactions between transparent institutions.

Prior approaches to achieving program equilibrium are often impractical for complex settings. In particular, the original program equilibrium papers [41, 33, 58, 65] are all based on the idea of comparing one's own source code with the opponent's. But in complex, asymmetric settings, it seems impossible to submit the same source code (without revealing parts of one's source code beforehand, which may not be in the players' interest). Some of my work therefore aims to achieve program equilibrium in more robust ways.

Secondly, program games generally allow for *many* equilibria. In fact, Rubinstein [58] and Tennenholtz [65] both give folk theorems for program equilibrium, which roughly matches the folk theorem for repeated games. (Every strategy profile whose utilities exceeds everyone's maximin payoff can be achieved in equilibrium.) Thus, while program games bring the potential for cooperation, they also come with the risk of unfair, socially suboptimal equilibria, or even

[2] I give an annotated bibliography on program equilibrium at `https://www.andrew.cmu.edu/user/coesterh/AnnotatedProgEqBibliography.html`.

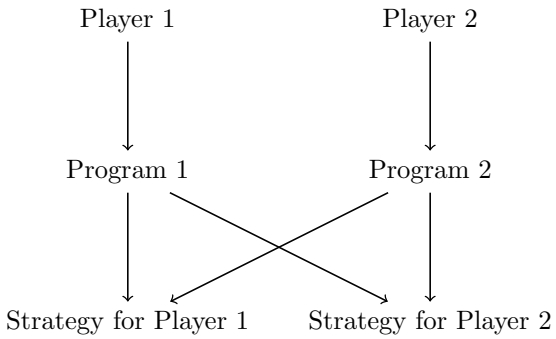

Figure 2: A graphical representation of a two-player program game. Player i's strategy in the underlying game is obtained by running Player i's program with Player $-i$'s program as input.

failure to coordinate on an equilibrium. Some of my work therefore investigates what equilibria can be achieved under various restrictions on the programs.

Key published papers

- Caspar Oesterheld. Robust program equilibrium. *Theory and Decision*, 86:143–159, 2019. DOI 10.1007/s11238-018-9679-3
- Caspar Oesterheld, Johannes Treutlein, Roger B Grosse, Vincent Conitzer, and Jakob Foerster. Similarity-based cooperative equilibrium. *Advances in Neural Information Processing Systems*, 36, 2024
  - Note that this paper also appears in the list in Section 3.3.
- Emery Cooper, Caspar Oesterheld, and Vincent Conitzer. Characterising simulation-based program equilibria. In *Proceedings of the Thirty-Ninth Annual AAAI Conference on Artificial Intelligence*, 2025

3.3 (Learning in) "Newcomb-like" decision problems

Consider Newcomb's problem [43], a well-known problem in (the more philosophical branches of) decision theory:

> Suppose a being in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, etc.) You know that this being has often correctly predicted your choices in the past (and has never, so far as you know, made an incorrect prediction about your choices), and further- more you know that this being has often correctly predicted the choices of other people, many of whom are similar to you, in the

> particular situation to be described below. [...] [A]ll this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct.
>
> There are two boxes, (B1) and (B2). (B1) contains $1,000. (B2) contains either $1,000,000 ($M), or nothing. What the content of (B2) depends upon will be described in a moment. You have a choice between two actions: (1) taking what is in both boxes; (2) taking only what is in the second box.
>
> Furthermore, and you know this, the being knows that you know this, and so on: (I) If the being predicts you will take what is in both boxes, he does not put the $M in the second box. (II) If the being predicts you will take only what is in the second box, he does put the $M in the second box. The situation is as follows. First the being makes its prediction. Then it puts the $M in the second box, or does not, depending upon what it has predicted. Then you make your choice. What do you do?

Decision theorists are famously divided over what the rational response to this problem is [8]. Supporters of so-called *evidential decision theory* argue that one should take only the opaque box, since doing so provides evidence that the box contains the million dollars. Conversely, supporters of so-called *causal decision theory* argue that one should take both boxes, because one's decision doesn't causally affect the contents of the opaque box, but is guaranteed to earn you an extra $1,000.[3]

Newcomb's problem itself is non-strategic, since the predictor is not a strategic agent. (We can easily make it strategic, though, by assuming the predictor's goal is simply to "match" the box contents with your eventual choice.) Nonetheless, Newcomb's problem is closely related to foundational questions in game theory. After all, game theory is all about situations in which agents have to make a decision under the knowledge that other players are predicting its decisions. Newcomb's problem poses a simple normative question about this type of situation.

While the foundational relevance of Newcomb's problem to game theory seems hard to argue against, the *practical* relevance to everyday *human* decision making is much less clear. It's hard to think of a current-day interaction in which (it's sufficiently clear that) a human decision maker's decisions are predictable in the way required by Newcomb's problem. Consequently, Newcomb's problem has received relatively little attention outside of philosophy.

For software agents, on the other hand, we might expect predictability to be the norm. After all, an agent's decision making algorithm could simply be made publicly available. (Compare Section 3.2.)

Additionally, software agents may often face copies of themselves and, for example, the Prisoner's Dilemma against a copy is a Newcomb-like problem

[3] Besides causal and evidential decision theory, some other theories have also been proposed [39].

[9, 40]. EDT recommends cooperation and CDT recommends defection. More broadly, EDT might allow for more cooperative, socially desirable behavior. For example, when ChatGPT decides whether to help Alice fill out a long-shot application to receive tax filing assistance from a non-profit, ChatGPT might take into account that if it helps, other instances of ChatGPT are more likely to analogously help with, say, long-shot applications. Thus, to decide whether to help, ChatGPT has to consider the consequences of a society-wide increase in long-shot applications to non-profits.[4] Plausibly these effects are harmful enough that even under alignment with the specific user, refusal is the correct response.

Besides increasing the relevance of reasoning about Newcomb-like problems, the AI perspective also often makes these situations less controversial. For instance, imagine that I build a software agent. The software agent then faces Newcomb's problem. (The predictor reads the source code that I write and makes a prediction based on that.) Then evidential and causal decision theorists agree that I should build an agent that *one*-boxes. After all, I (in contrast to my software agent) can causally influence the contents of the opaque box.

I've spent quite a bit of effort in my doctoral research on the decision theory of Newcomb-like problems (see a list of papers below). One of my papers directly contributes to the CDT versus EDT papers [48]. All my other papers specifically analyze how AI architectures and especially *learning algorithms* relate to reasoning about Newcomb-like problems.

Published papers

- Caspar Oesterheld and Vincent Conitzer. Extracting money from causal decision theorists. *The Philosophical Quarterly*, 71(4):pqaa086, 2021
- Caspar Oesterheld. Approval-directed agency and the decision theory of Newcomb-like problems. *Synthese*, 198(Suppl 27):6491–6504, 2021
- Caspar Oesterheld, Abram Demski, and Vincent Conitzer. A theory of bounded inductive rationality. In Rineke Verbrugge, editor, *Proceedings Nineteenth conference on Theoretical Aspects of Rationality and Knowledge*. arXiv, 2023
- James Bell*, Linda Linsefors*, Caspar Oesterheld*, and Joar Skalse*. Reinforcement learning in newcomblike environments. *Advances in Neural Information Processing Systems*, 34:22146–22157, 2021
- Caspar Oesterheld, Johannes Treutlein, Roger B Grosse, Vincent Conitzer, and Jakob Foerster. Similarity-based cooperative equilibrium. *Advances in Neural Information Processing Systems*, 36, 2024

  – Note that this paper also appears in the list in Section 3.2.

[4] For a discussion of this risk from the use of LLMs (the risk from removing mental-effort hurdles), see, e.g., Wojtowicz and DeDeo [72].

4 Current and planned work

4.1 Program equilibrium

- "Mediated versus program equilibrium": It turns out that program equilibrium is closely related to a concept called *mediated equilibrium* [42] (compare [35] for a setting that I'd view as being in between program and mediated equilibrium). Very roughly, a mediator in this context[5] is an entity that a contract on the table that the players can simultaneously (independently) decide to sign or not sign. Everyone who signs the contract is bound by what is specified by the contract, but the actions mandated by the contract can depend on the set of players that sign the contract. A good real-world example of such a contract is the *National Popular Vote Interstate Compact (NPVIC)*[6], an agreement between states in the US to give their electoral college votes to the winner of the popular vote *if more than half the popular vote's worth of states sign the NPVIC*, and to vote in the traditional way (in most states for the winner of the specific state) otherwise. In the Prisoner's Dilemma, the following contract turns cooperation into an equilibrium: Both players must cooperate if the contract is signed by everyone; otherwise whoever signs this contract must defect.

  One desirable property of mediated equilibrium is that it allows only a specific equilibrium (as given by the mediator). Thus, it avoids the equilibrium selection problem that plagues other notions of commitment. Of course, this benefit comes (at least in the basic story described above) at a cost of centralization.

  It turns out that mediators can allow the same set of equilibria that program equilibrium allows. (I have an independent proof of this, but it also follows from existing work, e.g., combining the folk theorems of Tennenholtz [66] and Kalai et al. [35].)

  Following Monderer and Tennenholtz [42], we have also investigated solution concepts based on multi-player deviations (whereas Nash equilibrium, and thus program and mediated equilibrium). Here we find differences between the program game and the mediator concepts: because deviations are somewhat transparent in the program game setting, they can be punished more effectively.

  My goal is to write these results up into a (conference-quality) paper to be included in my dissertation.

- "Sequential program equilibrium": As far as I know, prior work on program games and mediators has only considered *simultaneous* play. What

[5] Others have used the term "mediator" to refer to completely different game-theoretic concepts. The meaning used here also differed from how I'd understand the term "mediator" in common English.

[6] `https://en.wikipedia.org/wiki/National_Popular_Vote_Interstate_Compact`

happens if the players get to choose programs (or contracts) *sequentially*? For instance, imagine Player 1 can propose and sign a contract (which governs what Player 1 and anyone else who signs the contract will choose); then Player 2 gets to see the contract and chooses whether to sign the contract or propose a new contract independently; and finally Player 3 gets to see all previously proposed contracts and either signs a contract or chooses independently.

Compared to mediated and program equilibrium, the dynamics governing what payoffs will be achieved in this sequential setting are much more complicated. In fact, I have a (fairly complicated) proof sketch (worked out as a lead contributor in a project with Vincent Conitzer and Jiayuan Liu) that shows that it is NP-hard to determine what happens, even in the three player case.

I hope to write this result up into a paper.

4.2 Connecting Newcomb-like problems to self-locating beliefs

It turns out that there is a close connection between Newcomb-like decision problems and so-called *self-locating beliefs* (a.k.a. anthropics or beliefs under imperfect recall) [57, 23, 7].

- First, it turns out that decision-theoretic questions come up when making decisions in decision problems of imperfect recall. For instance, consider a variant of the Sleeping Beauty problem [23] in which a decision maker is offered bets on the outcome of the coin flip. Should Beauty take into account that accepting the bet provides that she also accepts the bet on other awakenings? Evidential decision theory gives a positive answer; causal decision theory gives a negative answer. This has been pointed out before [36, 14, 1, 10, 62].

  Most closely following prior work by Briggs [10], I have a working paper with Vince, showing that to make good decisions, we need to match EDT with the so-called (double) halfer's position and CDT with the so-called thirder's position.

  Working draft: Caspar Oesterheld and Vincent Conitzer. Can de se choice be ex ante reasonable in games of imperfect recall? `https://www.andrew.cmu.edu/user/coesterh/DeSeVsExAnte.pdf`

- Conversely, the self-locating beliefs perspective might come up when facing Newcomb-like decision problems. For instance, in Newcomb's problem, should I believe that I am in a simulation in the predictor's mind? If so, CDT, too, might recommend one-boxing in order to cause the opaque box to be filled.

Working draft: Emery Cooper, Caspar Oesterheld, and Vincent Conitzer. Can CDT rationalise the ex ante optimal policy via modified anthropics? *arXiv preprint arXiv:2411.04462*, 2024

4.3 Safe Pareto improvements

- With Vince Conitzer, I am working on a draft, which takes a relatively abstract perspective on safe Pareto improvements (whereas the original safe Pareto improvements paper [49] focuses on a fairly specific class of interventions on games).

  Working paper: Caspar Oesterheld and Vincent Conitzer. Choosing what game to play with no regrets or controversies – results on inferring safe (Pareto) improvements in binary constraint structures, 2024. URL `https://www.andrew.cmu.edu/user/coesterh/SPIxBCS.pdf`[7]

- In equal co-authorshop with Nathaniel Sauerberg, I am working on a project in which we consider safe Pareto improvements under what we call *ex post verifiable* commitments on (normal-form) games. For instance, we ask: if a player credibly commits against taking a particular action, is the resulting game a safe Pareto improvement on the original game (the game prior to the commitment). We give a mix of characterizations, efficient algorithms and hardness results.

  We plan to have a full version of the paper by the time of my graduation.

  Working draft: Nathaniel Sauerberg* and Caspar Oesterheld*. Promises made, promises kept: Safe Pareto improvements via ex post verifiable commitments

4.4 Implementing theoretical ideas with language models

Over the past two to three years, language models have emerged as a leading paradigm in AI. To assess the applicability of my theoretical work, I've started exploring their implementation in language models. Throughout my PhD my work has been relatively abstract, in part because I want it to be broadly applicable. Now seems like a good time to test whether I have achieved this goal. I discuss two projects that have progressed relatively far below, and I will likely work on additional projects as well.

- In joint work with Maxime Riché and Filip Sondej, as well as Jesse Clifton and Vince, I have worked on a project on implementing safe Pareto improvements (Section 3.1) in language models, in particular so-called *surrogate goals*.

  Working paper: Caspar Oesterheld*, Maxime Riché*, Filip Sondej*, Jesse Clifton, and Vincent Conitzer. Implementing surrogate goals for safer

[7] My blog post for the PhD program's blog post requirements (available at `https://hackmd.io/PIbBmbx_QWK52cCQsyax0w`) takes a similar high-level perspective as this paper.

bargaining in LLM-based agents. URL https://www.andrew.cmu.edu/user/coesterh/LLMxSG.pdf

- As discussed in Section 3.3, reasoning in particular ways about Newcomb-like problems may enable greater cooperation. In joint work with Emery Cooper, Chi Nguyen, Miles Kodama, and Ethan Perez, I've developed a large, diverse, high-quality, hand-generated dataset of natural-language questions about Newcomb-like problems, and analyze responses by a large number of current language models.

  Working paper: Caspar Oesterheld, Emery Cooper, Miles Kodama, Linh Chi Nguyen, and Ethan Perez. A dataset of questions on decision-theoretic reasoning in Newcomb-like problems. URL https://arxiv.org/abs/2411.10588

References

[1] Stuart Armstrong. Anthropic decision theory, 2011. URL https://arxiv.org/abs/1110.6437.

[2] Computer Science Department at Carnegie Mellon University. The computer science phd program at carnegie mellon university, 9 2024. URL https://www.csd.cs.cmu.edu/sites/default/files/2024-09/CSD-PhD-Handbook-2024-25_0.pdf.

[3] Mihaly Barasz, Paul Christiano, Benja Fallenstein, Marcello Herreshoff, Patrick LaVictoire, and Eliezer Yudkowsky. Robust cooperation in the prisoner's dilemma: Program equilibrium via provability logic, 1 2014. URL https://arxiv.org/abs/1401.5577.

[4] Tobias Baumann. Using surrogate goals to deflect threats, 12 2017. URL https://s-risks.org/using-surrogate-goals-to-deflect-threats/.

[5] James Bell*, Linda Linsefors*, Caspar Oesterheld*, and Joar Skalse*. Reinforcement learning in newcomblike environments. *Advances in Neural Information Processing Systems*, 34:22146–22157, 2021.

[6] Joyce E. Berg and Thomas A. Rietz. Prediction markets as decision support systems. *Information Systems Frontiers*, 5(1):79–93, 2003. doi: 10.1023/A:1022002107255.

[7] Nick Bostrom. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Studies in Philosophy. Routledge, 2010.

[8] David Bourget and David J Chalmers. What do philosophers believe? *Philosophical studies*, 170:465–500, 2014.

[9] Steven J. Brams. Newcomb's problem and prisoners' dilemma. *The Journal of Conflict Resolution*, 19(4):596–612, 12 1975.

[10] Rachael Briggs. Putting a value on beauty. In *Oxford Studies in Epistemology*, volume 3, pages 3–24. Oxford University Press, 2010. URL http://joelvelasco.net/teaching/3865/briggs10-puttingavalueonbeauty.pdf.

[11] Yiling Chen, Ian A. Kash, Mike Ruberry, and Victor Shnayder. Eliciting predictions and recommendations for decision making. In *ACM Transactions on Economics and Computation*, volume 2, chapter 6. 6 2014. URL http://yiling.seas.harvard.edu/wp-content/uploads/a6-chen.pdf.

[12] Phillip JK Christoffersen, Andreas A Haupt, and Dylan Hadfield-Menell. Get it in writing: Formal contracts mitigate social dilemmas in multi-agent rl. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 448–456, 2023.

[13] Jesse Clifton. Cooperation, conflict, and transformative artificial intelligence: A research agenda. *Effective Altruism Foundation, March*, 4, 2020.

[14] Vincent Conitzer. A dutch book against sleeping beauties who are evidential decision theorists. *Synthese*, 192(9):2887–2899, 10 2015. doi: 10.1007/s11229-015-0691-7.

[15] Vincent Conitzer. Designing preferences, beliefs, and identities for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9755–9759, 2019.

[16] Vincent Conitzer and Caspar Oesterheld. Foundations of cooperative AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15359–15367, Sep. 2023. doi: 10.1609/aaai.v37i13.26791. URL `https://ojs.aaai.org/index.php/AAAI/article/view/26791`.

[17] Emery Cooper, Caspar Oesterheld, and Vincent Conitzer. Can CDT rationalise the ex ante optimal policy via modified anthropics? *arXiv preprint arXiv:2411.04462*, 2024.

[18] Emery Cooper, Caspar Oesterheld, and Vincent Conitzer. Characterising simulation-based program equilibria. In *Proceedings of the Thirty-Ninth Annual AAAI Conference on Artificial Intelligence*, 2025.

[19] Andrew Critch. A parametric, resource-bounded generalization of löb's theorem, and a robust cooperation criterion for open-source game theory. *Journal of Symbolic Logic*, 84(4):1368–1381, 12 2019. doi: 10.1017/jsl.2017.42.

[20] Andrew Critch and David Krueger. AI research considerations for human existential safety (arches). *arXiv preprint arXiv:2006.04948*, 2020.

[21] Andrew Critch, Michael Dennis, and Stuart Russell. Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory. *arXiv preprint arXiv:2208.07006*, 2022.

[22] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*, 2020.

[23] Adam Elga. Self-locating belief and the sleeping beauty problem. *Analysis*, 60(2):143–147, 2000.

[24] Scott Emmons, Caspar Oesterheld, Andrew Critch, Vincent Conitzer, and Stuart Russell. For learning in symmetric teams, local optima are global nash equilibria. In *International Conference on Machine Learning*, pages 5924–5943. PMLR, 2022.

[25] Scott Emmons*, Caspar Oesterheld*, Vincent Conitzer, and Stuart Russell. Observation interference in partially observable assistance games. *arXiv preprint arXiv:2412.17797*, 2024.

[26] Ivan Geffner, Caspar Oesterheld, and Vincent Conitzer. Maximizing social welfare with side payments. Draft submitted to EC '25.

[27] Anshul Gupta and Sven Schewe. It pays to pay in bi-matrix games: A rational explanation for bribery. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, page 1361–1369, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450334136.

[28] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced AI. Technical Report 1, Cooperative AI Foundation, 2025.

[29] Robin Hanson. Decision markets. *IEEE Intelligent Systems*, 14(3):16–19, 1999. URL `http://mason.gmu.edu/~rhanson/decisionmarkets.pdf`.

[30] Robin Hanson. Decision markets. In Alexander Tabarrok, editor, *Entrepreneurial Economics – Bright Ideas from the Dismal Science*, chapter 5, pages 79–85. Oxford University Press, 2002.

[31] Robin Hanson. Decision markets for policy advice. In Alan S. Gerber and Eric M. Patashnik, editors, *Promoting the General Welfare: New Perspectives on Government Performance*, pages 151–173. Brookings Institution Press, 11 2006.

[32] Robin Hanson. Shall we vote on values, but bet on beliefs? *Journal of Political Philosophy*, 21(2), 2013. doi: 10.1111/jopp.12008.

[33] J. V. Howard. Cooperation in the prisoner's dilemma. *Theory and Decision*, 24:203–213, 5 1988. doi: 10.1007/BF00148954.

[34] Matthew O Jackson and Simon Wilkie. Endogenous games and mechanisms: Side payments among players. *The Review of Economic Studies*, 72 (2):543–566, 2005.

[35] Adam Tauman Kalai, Ehud Kalai, Ehud Lehrer, and Dov Samet. A commitment folk theorem. *Games and Economic Behavior*, 69(1):127–137, 2010.

[36] Theodore Korzukhin. A dutch book for cdt thirders. *Synthese*, 2020. doi: 10.1007/s11229-020-02841-7.

[37] Vojtech Kovarik, Caspar Oesterheld, and Vincent Conitzer. Game theory with simulation of other players. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2800–2807, 2023.

[38] Vojtech Kovarik, Caspar Oesterheld, and Vincent Conitzer. Recursive joint simulation in games. *arXiv preprint arXiv:2402.08128*, 2024.

[39] Benjamin A Levinstein and Nate Soares. Cheating death in damascus. *The Journal of Philosophy*, 117(5):237–266, 2020.

[40] David Lewis. Prisoners' dilemma is a Newcomb problem. *Philosophy & Public Affairs*, 8(3):235–240, 1979.

[41] R. Preston McAfee. Effective computability in economic decisions. 5 1984. URL `https://www.mcafee.cc/Papers/PDF/EffectiveComputability.pdf`.

[42] Dov Monderer and Moshe Tennenholtz. Strong mediated equilibrium. 173 (1):180–195, 2009.

[43] Robert Nozick. Newcomb's problem and two principles of choice. In *Essays in honor of Carl G. Hempel: A tribute on the occasion of his sixty-fifth birthday*, pages 114–146. Springer, 1969.

[44] Caspar Oesterheld. Robust program equilibrium. *Theory and Decision*, 86: 143–159, 2019. DOI 10.1007/s11238-018-9679-3.

[45] Caspar Oesterheld. Approval-directed agency and the decision theory of Newcomb-like problems. *Synthese*, 198(Suppl 27):6491–6504, 2021.

[46] Caspar Oesterheld and Vincent Conitzer. Can de se choice be ex ante reasonable in games of imperfect recall? `https://www.andrew.cmu.edu/user/coesterh/DeSeVsExAnte.pdf`.

[47] Caspar Oesterheld and Vincent Conitzer. Decision scoring rules. In *Web and Internet Economics: 16th International Conference, WINE 2020, Beijing, China, December 7–11, 2020, Proceedings*, volume 12495, page 468. Springer Nature, 2020.

[48] Caspar Oesterheld and Vincent Conitzer. Extracting money from causal decision theorists. *The Philosophical Quarterly*, 71(4):pqaa086, 2021.

[49] Caspar Oesterheld and Vincent Conitzer. Safe Pareto improvements for delegated game playing. *Autonomous Agents and Multi-Agent Systems*, 36 (2), 2022. Article number 46.

[50] Caspar Oesterheld and Vincent Conitzer. Choosing what game to play with no regrets or controversies – results on inferring safe (Pareto) improvements in binary constraint structures, 2024. URL `https://www.andrew.cmu.edu/user/coesterh/SPIxBCS.pdf`.

[51] Caspar Oesterheld, Emery Cooper, Miles Kodama, Linh Chi Nguyen, and Ethan Perez. A dataset of questions on decision-theoretic reasoning in Newcomb-like problems. URL `https://arxiv.org/abs/2411.10588`.

[52] Caspar Oesterheld*, Maxime Riché*, Filip Sondej*, Jesse Clifton, and Vincent Conitzer. Implementing surrogate goals for safer bargaining in LLM-based agents. URL `https://www.andrew.cmu.edu/user/coesterh/LLMxSG.pdf`.

[53] Caspar Oesterheld, Abram Demski, and Vincent Conitzer. A theory of bounded inductive rationality. In Rineke Verbrugge, editor, *Proceedings Nineteenth conference on Theoretical Aspects of Rationality and Knowledge*. arXiv, 2023.

[54] Caspar Oesterheld, Johannes Treutlein, Emery Cooper, and Rubi Hudson. Incentivizing honest performative predictions with proper scoring rules. In *Uncertainty in Artificial Intelligence*, pages 1564–1574. PMLR, 2023.

[55] Caspar Oesterheld, Johannes Treutlein, Roger B Grosse, Vincent Conitzer, and Jakob Foerster. Similarity-based cooperative equilibrium. *Advances in Neural Information Processing Systems*, 36, 2024.

[56] Abraham Othman and Tuomas Sandholm. Decision rules and decision markets. In *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010), van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada*, pages 625–632. 2010.

[57] Michele Piccione and Ariel Rubinstein. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior*, 20:3–24, 1997. doi: 10.1006/game.1997.0536.

[58] Ariel Rubinstein. *Modeling Bounded Rationality*. Zeuthen Lecture Book Series. The MIT Press, 1998.

[59] Stuart Russell. *Human Compatible. AI and the Problem of Control*. Penguin, 2019.

[60] Nathaniel Sauerberg* and Caspar Oesterheld*. Promises made, promises kept: Safe Pareto improvements via ex post verifiable commitments.

[61] Nathaniel Sauerberg and Caspar Oesterheld. Computing optimal commitments to strategies and outcome-conditional utility transfers. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1654–1663, 2024.

[62] Wolfgang Schwarz. Lost memories and useless coins: revisiting the absentminded driver. *Synthese*, 192:3011–3036, 2015. doi: 10.1007/s11229-015-0699-z.

[63] Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. Benefits of assistance over reward learning.

[64] Nate Soares and Benya Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. *The technological singularity: Managing the journey*, pages 103–125, 2017.

[65] Moshe Tennenholtz. Program equilibrium. 49:363–373, 2004.

[66] Moshe Tennenholtz. Program equilibrium. *Games and Economic Behavior*, 49(2):363–373, 2004. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2004.02.002. URL https://www.sciencedirect.com/science/article/pii/S0899825604000314.

[67] Emanuel Tewolde, Caspar Oesterheld, Vincent Conitzer, and Paul W Goldberg. The computational complexity of single-player imperfect-recall games. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2878–2887, 2023.

[68] Emanuel Tewolde, Brian Hu Zhang, Caspar Oesterheld, Manolis Zampetakis, Tuomas Sandholm, Paul Goldberg, and Vincent Conitzer. Imperfect-recall games: equilibrium concepts and their complexity. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 2994–3004, 2024.

[69] Emanuel Tewolde, Brian Hu Zhang, Caspar Oesterheld, Tuomas Sandholm, and Vincent Conitzer. Computing game symmetries and equilibria that respect them. In *Proceedings of the Thirty-Ninth Annual AAAI Conference on Artificial Intelligence*, 2025.

[70] Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. A new formalism, method and open issues for zero-shot coordination. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10413–10423. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/treutlein21a.html.

[71] Richard Willis and Michael Luck. Resolving social dilemmas through reward transfer commitments. In *Adaptive and Learning Agents Workshop*, 2023. URL https://kclpure.kcl.ac.uk/ws/portalfiles/portal/207013371/ALA2023_paper_65.pdf.

[72] Zachary Wojtowicz and Simon DeDeo. Undermining mental proof: How ai can make cooperation harder by making thinking easier. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25),*, 2025. URL https://arxiv.org/pdf/2407.14452.

A Other work and how it fits in

I here list a few other high-level topics that I've worked on during my PhD. Generally these works are less central to the high-level objectives of the dissertation. They also often involve other co-authors more heavily, making inclusion in the dissertation tricky. Finally, I don't want to write (and I assume my committee members don't want to read) an overly long dissertation.

A.1 Equilibrium selection in (symmetric) common-payoff games

Even in *common-payoff* games (i.e., games in which all players have the same utility), equilibrium selection problems can arise. Common-payoff games and equilibrium selection therein has seen a recent surge in interest as a model of interactions between a human and an AI *assistant* [59, 63]. Even two-player common-payoff games are surprisingly rich in some ways.

Besides equilibrium selection, there is another, perhaps more surprising, connection to my work: There is a close connection between choice under imperfect recall (see [57, 23]) and playing a symmetric team game with symmetry constraints.

Besides my aforementioned work on imperfect recall and its connection to EDT/CDT, I've been involved in a few papers relating to equilibrium selection in common-payoff team games, with a special focus on games with symmetries.

Finally, there's a more technical connection to my work: identifying symmetries in a game is closely related to identifying *isomorphisms* between games, which in turn is a key to identifying safe Pareto improvements.

My work on (symmetric) common-payoff games

- Scott Emmons*, Caspar Oesterheld*, Vincent Conitzer, and Stuart Russell. Observation interference in partially observable assistance games. *arXiv preprint arXiv:2412.17797*, 2024
- Symmetries:
  - Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. A new formalism, method and open issues for zero-shot coordination. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10413–10423. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/treutlein21a.html`
  - Scott Emmons, Caspar Oesterheld, Andrew Critch, Vincent Conitzer, and Stuart Russell. For learning in symmetric teams, local optima are global nash equilibria. In *International Conference on Machine Learning*, pages 5924–5943. PMLR, 2022

– Emanuel Tewolde, Brian Hu Zhang, Caspar Oesterheld, Tuomas Sandholm, and Vincent Conitzer. Computing game symmetries and equilibria that respect them. In *Proceedings of the Thirty-Ninth Annual AAAI Conference on Artificial Intelligence*, 2025

A.2 Commitment to payments

One high-level idea for addressing the equilibrium selection problem is to identify mechanisms that (a) introduce cooperative equilibria, but (b) don't also introduce lots of undesirable equilibria.

A natural candidate for such a mechanism is to allow players to commit to outcome/behavior-conditional utility transfers (e.g., monetary payoffs). After all, real-world cooperation (e.g., the cooperation between an employer and employee, or between a customer and a merchant) often works by having one party commit to pay another if the other party behaves in a desirable way. Furthermore, it appears on first sight that offering to pay someone can do relatively little harm. In the worst case, one might think, one can simply ignore such offers.

An optimistic view of offers to payments has motivated a number of papers (e.g., [34, 12, 27, 71]). I've been involved in some work on "committing to payments" (see below). However, my plan is to not feature this work much in my dissertation. For one, I've so far only worked on this as a non-primary author. Second, during our own work we found that committing to payments in general normal-form games is a more double-edged sword than one might expect (cf. [34]). (That said, there are various directions toward more positive results, e.g., [71, 26].)

My work on committing to payments

- Nathaniel Sauerberg and Caspar Oesterheld. Computing optimal commitments to strategies and outcome-conditional utility transfers. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1654–1663, 2024
- Ivan Geffner, Caspar Oesterheld, and Vincent Conitzer. Maximizing social welfare with side payments. Draft submitted to EC '25

A.3 Predictions, decisions and counterfactuals

One perspective on Newcomb-like problems (as discussed in Section 3.3) is that predictions about counterfactual (i.e., untaken) actions are problematic. The one-boxer finds the opaque box full and believes that were he to two-box instead, the opaque box would be empty. The two-boxer expects the opaque box empty, but believes that were she to one-box, the opaque box would still be empty. So the one- and two-boxer might agree on what will in fact happen. But they might disagree about what would happen if they were to take some different action. Of course, the disagreement between these counterfactual beliefs is never tested.

In some sense, the belief that the other action is bad causes this belief to remain untested.

This phenomenon – predictions about the consequences of different decisions influence which – can arise in other contexts as well. For instance, imagine that we use a prediction market to predict the consequences of different actions, and then use those predictions to make a decision [29, 30, 6, 31, 32]. Othman and Sandholm [56] show that predictors are not necessarily incentivized to report their predictions truthfully. A few different solutions to this problem has been proposed [11]. I've proposed my own solution (which avoids eliciting beliefs about counterfactuals altogether) [47], which has inspired my later work on learning in Newcomb-like settings that allows for cooperation in a Prisoner's Dilemma against a similar opponent [53].

My work on predictions, decisions and counterfactuals

- Caspar Oesterheld and Vincent Conitzer. Decision scoring rules. In *Web and Internet Economics: 16th International Conference, WINE 2020, Beijing, China, December 7–11, 2020, Proceedings*, volume 12495, page 468. Springer Nature, 2020

- Caspar Oesterheld, Johannes Treutlein, Emery Cooper, and Rubi Hudson. Incentivizing honest performative predictions with proper scoring rules. In *Uncertainty in Artificial Intelligence*, pages 1564–1574. PMLR, 2023

A.4 Games with simulations of other players

My work on program equilibrium has focused on programs that simulate each other, e.g. in the Prisoner's Dilemma: "with probability ϵ, cooperate; with the remaining probability, simulate the opponent and copy whatever action they play" [44]. I've also been involved in a line of work led by Vojta Kovarik on what happens if (instead of the full program equilibrium framework) we specifically give the players the ability to simulate each other.

While this line of work is both closely related to the program equilibrium work and to the goal of achieving cooperation, it will not feature centrally in the dissertation, given that Vojta has been the lead author on this line of work.

A.4.1 My work on games with simulations of other players

- Vojtech Kovarik, Caspar Oesterheld, and Vincent Conitzer. Recursive joint simulation in games. *arXiv preprint arXiv:2402.08128*, 2024

- Vojtech Kovarik, Caspar Oesterheld, and Vincent Conitzer. Game theory with simulation of other players. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2800–2807, 2023

B Biographical information

My CV can be found at `https://www.andrew.cmu.edu/user/coesterh/cv.pdf` and my Google Scholar profile is at `https://scholar.google.com/citations?user=xeEcRjkAAAAJ`.

Most of my work is technical (theoretical computer science and game theory). I have a strong technical background in mathematics. When I was in (the German equivalent of) middle and high school, I took about a year's worth of coursework at the University of Hamburg (incl. Analysis I–III, Linear Algebra I and II, and Discrete Mathematics). I also made it to the national round of the Math Olympiad. Due to my interest in AI, I proceeded to obtain an undergraduate degree in computer science, focusing my coursework on AI and theoretical computer science (taking, among others, Math for CS, Theoretical CS I and II, Algorithms on Graphs, Logic, Practical CS I–III, AI, Cognitive Systems, Formal Modeling, Petri Nets, and ML). I similarly focused my coursework at Duke and CMU (Computational Microeconomics, Algorithms for Decision Making at Scale, Information Theory, ML and Game Theory, AI). I've published 16 papers at major AI and multi-agent systems conferences (AAAI, AAMAS, IJCAI, NeurIPS, ICML, TARK, WINE, UAI), most of which are closely related to the focus of my dissertation.

Some of my work touches on more philosophical issues in decision theory. Unfortunately, I have not enjoyed any formal, post-secondary education in philosophy. Nonetheless, I've been able to publish some of my work in highly regarded, peer-reviewed philosophical venues. In particular, I have two publications at *Synthese*, a leading philosophy journal[8] and perhaps *the* leading philosophy journal that commonly publishes work in mathematical philosophy, as well as a paper in the highly regarded *The Philosophical Quarterly*[9].

I've been awarded with a PhD fellowship from the Future of Life Institute.

C Instructions for the thesis proposal document

From the CSD PhD Student handbook [2, Sect. 10.3]: "T"he student submits a written proposal to the faculty. The student also orally presents the thesis proposal to interested faculty and students in a public colloquium.

A thesis proposal should:

- Explain the basic idea of the thesis topic (e.g., the problem to be solved and the approach to solving it).
- Argue why that topic is interesting (e.g., what contributions to the field would be made in carrying out the proposed work).
- State what kind of results have already been obtained and what further results are expected.

[8] A 2022 survey by Brian Leiter ranks *Synthese* as the 11th-best philosophy jorunal.
[9] Leiter's ranking puts *The Philosophical Quarterly* at 9th.

- Argue that these results are obtainable within a reasonable amount of time.
- Demonstrate the student's personal qualifications for doing the proposed work.

The main purpose of the thesis proposal is to convince the faculty that the chosen thesis topic is significant and that the student's approach has a reasonable chance of success. [...] A thesis proposal should be short, about 15–20 pages.

A thesis proposal should not be:

- A dry run for the thesis
- A summary or abstract of the thesis
- The first chapter or part of the thesis
- A technical report
- A survey of the field
- An annotated bibliography

Any included list of references or bibliography should serve the purpose of supporting the assessment of the state of the art and the student's personal qualifications.