Foundations of Cooperative AI

Vincent Conitzer[1], **Caspar Oesterheld**[1]

[1]Computer Science Department
Carnegie Mellon University
conitzer@cs.cmu.edu, oesterheld@cmu.edu

Abstract

AI systems can interact in unexpected ways, sometimes with disastrous consequences. As AI gets to control more of our world, these interactions will become more common and have higher stakes. As AI becomes more advanced, these interactions will become more sophisticated, and game theory will provide the tools for analyzing these interactions. However, AI agents are in some ways unlike the agents traditionally studied in game theory, introducing new challenges as well as opportunities. We propose a research agenda to develop the game theory of highly advanced AI agents, with a focus on achieving cooperation.

Introduction

AI safety is a nascent research area aiming to prevent harmful unintended behavior in advanced AI systems (Amodei et al. 2016; Russell 2019). One way in which AI systems can fail to be safe is through unexpected interactions with each other. A well-known example of disastrous algorithmic interaction is the 2010 "flash crash," in which the Dow Jones Industrial Average dropped by about 9% in a very short period of time, and in which high-frequency traders (who use algorithmic trading) played a major role (CFTC and SEC 2010). As AI systems control an ever growing part of our world, it stands to reason that they will run up against each other increasingly often, with the potential for disastrous interactions in many new domains, including ones in which this will have immediate significant physical impact on the world. (Consider, for example, autonomous vehicles, use of AI in the electrical grid, or military applications of AI.) On the other hand, AI may also provide help in addressing challenging collective action problems that humans already face, such as climate change and other environmental problems, and nuclear disarmament and the prevention of wars. It may do so, for example, by providing better monitoring and transparency (Russell, Vaidya, and Bras 2010).

These issues lead us to want to design *cooperative AI* (Dafoe et al. 2020, 2021). One can take a variety of approaches to cooperative AI, including behavioral (i.e., by studying how humans cooperate with each other) and experimental ones. However, we argue that the *theoretical foundations* of cooperative AI are most naturally established with a game-theoretic approach.[1] Game theory provides the basic language and tools for reasoning about settings in which multiple agents each pursue their own objectives. However, we argue that to be able to model advanced AI systems, and the ways in which they may attain cooperative behavior, new work at the foundations of game theory is required.

One might argue that ideally, no game-theoretic phenomena should occur in the deployment of AI, if we are sufficiently careful to align its objectives with our values. However, we now give an example that shows that even if each *individual* agent is almost perfectly aligned with our (say, humanity's) objective, it is possible that the only *equilibrium* of the resulting game between the agents results in a terrible outcome. That is, aligning individual agents is not sufficient for the multiagent system *as a whole* to be well aligned with our objectives, due to game-theoretic phenomena. (Alternatively, the same game can be thought of as between two agents, each of which is *perfectly* aligned with a distinct subset of humanity, but the two subsets of humanity have slightly conflicting objectives.)

Consider the following setting. Two AI agents together provide a service. Each agent $i \in \{1, 2\}$ chooses a level of quality q_i with which to provide its part of the service. The overall quality of the service provided is $q = \min_{i \in \{1,2\}} q_i$. That is, the service is only as good as its weaker part. Let us say that our (humanity's) objective is q; we are not a player in the game. As for the agents, suppose that agent i's utility is $q + \mathbf{1}_{q_i < q_{-i}} \epsilon$, where $\mathbf{1}_{q_i < q_{-i}} = 1$ if q_i is lower than the other agent's choice of quality, $\mathbf{1}_{q_i < q_{-i}} = 1/2$ if the two agents choose the same quality, and $\mathbf{1}_{q_i < q_{-i}} = 0$ otherwise. That is, each agent cares almost exclusively about q, but has a slight incentive to be the one choosing the lower quality.

Table 1 shows an example of this game where $\epsilon = 4$ and qualities are integers from 0 to 7. Note that within each entry, the numbers are close to each other and to q (at most $\epsilon = 4$ apart), indicating that the agents are almost perfectly aligned;[2] also, of course, there are high-quality outcomes when both agents choose high numbers. Nevertheless, dis-

[1]Confusingly, while we take a game-theoretic approach to cooperative AI, what we are interested in is not *cooperative game theory* (though the latter may yet play a role in cooperative AI). Cooperative game theory concerns the formation and actions of coalitions of agents, under the assumption that some mechanism is available to these coalitions for enforcing agreements.

[2]It is in fact possible to make the alignment arbitrarily close to perfect by increasing the number of values of q to choose from.

	7	6	5	4	3	2	1	0
7	9,9	6,10	5,9	4,8	3,7	2,6	1,5	0,4
6	10,6	8,8	5,9	4,8	3,7	2,6	1,5	0,4
5	9,5	9,5	7,7	4,8	3,7	2,6	1,5	0,4
4	8,4	8,4	8,4	6,6	3,7	2,6	1,5	0,4
3	7,3	7,3	7,3	7,3	5,5	2,6	1,5	0,4
2	6,2	6,2	6,2	6,2	6,2	4,4	1,5	0,4
1	5,1	5,1	5,1	5,1	5,1	5,1	3,3	0,4
0	4,0	4,0	4,0	4,0	4,0	4,0	4,0	2,2

Table 1: The Traveler's Dilemma.

	Cooperate	Defect
Cooperate	$3, 3$	$0, 4$
Defect	$4, 0$	$1, 1$

Table 2: The Prisoner's Dilemma.

astrously, the only equilibrium is $q_1 = q_2 = 0$. (In fact, $q_i = 0$ is the only rationalizable strategy in this game.) Typical learning algorithms that only concern the learning agent's utility (e.g., no-regret learning, Q-learning with random exploration, gradient ascent) would converge to this equilibrium.

The game described is known in the literature as the *Traveler's Dilemma* (Basu 1994). For most of the paper, it will suffice to consider the simpler *Prisoner's Dilemma*, shown in Figure 2. As is well known, in the Prisoner's Dilemma, each agent has a *strictly dominant* strategy to defect, in spite of the fact that both agents are better off if both agents cooperate than if both agents defect.

To lay out the research agenda, in the remainder of this paper, we overview some possible ways in which AI agents could yet reach cooperation in such settings. They are roughly ordered to start with relatively traditional approaches from the game theory literature that make sense for human agents as well, to novel approaches that seem farfetched for human agents but may fit AI well. We close with a call to action, to integrate and expand on these research topics to realize the vision of cooperative AI.

Cooperation in repeated games

Perhaps the best-known way to achieve cooperation in games such as the Prisoner's Dilemma is by *repeating* the game. This, indeed, is many game theorists' go-to explanation for why, in our human world, we see more cooperation in these types of games than the straightforward analysis might suggest. The intuitive reasoning is that while we may often have a short-term incentive for taking advantage of another (defecting), in the longer term we are likely to run into the same person again, and defecting on the person now is likely to result in that person defecting on us in the future. The theory of repeated games formalizes this. Specifically, we consider infinitely repeated games, where a game is repeated over and over again and (say) the payoff in round t is discounted using a factor γ^t for some $\gamma < 1$. In such games, a *strategy* specifies what a player would do for every *history* of the game. Consider the following ("grim trigger") strategy for the Prisoner's Dilemma: if in the history so far there has never been any defection, then cooperate; otherwise, defect. Some reflection reveals that, for sufficiently large γ, both players playing the grim-trigger strategy is an equilibrium of the Prisoner's Dilemma, and one in which both players in fact cooperate forever.[3]

What else can be achieved in equilibria of repeated games? This question is answered by the *folk theorem* – really, a collection of theorems – characterizing which payoff profiles are possible to sustain in equilibria of repeated games, including ones other than the Prisoner's Dilemma. Generally they allow for a vast variety of equilibrium payoffs, restricted only by what payoffs are feasible in the game and the constraint that a player cannot be made worse off than she would be if everyone else conspired to make her miserable. In particular, because the second constraint only places *lower* bounds on players' utilities, these theorems generally allow high-welfare equilibria. Moreover, intriguingly, folk theorems can facilitate equilibrium computation, which, at least for two players, is easier in repeated games than in one-shot games (Littman and Stone 2005; Andersen and Conitzer 2013). (For three or more players, the story becomes more complicated (Borgs et al. 2010; Kontogiannis and Spirakis 2008).)

Is this, then, perhaps a completely satisfactory solution to the problem of cooperation? Unfortunately, it seems that that is not the case. First of all, some games are *not* repeated. A nuclear war is fought only once. (We will turn to disarmament in the next section.) Even if the game is repeated, the vast multiplicity of equilibria allowed by the folk theorem creates a problem of *equilibrium selection* – as a player in a repeated game, which of the many equilibria should you play? We will return to this topic in a later section. Even assuming the equilibrium selection problem can be resolved, repeated games allow for cooperative equilibria only under certain conditions. The lower the discount factor and the lower the benefit of cooperation, the less cooperation can be sustained. Also, worrisomely, there is often only a thin line dividing total cooperation from total defection. Moon and Conitzer (2015) study a model in which players interact on a social network, so that when player 1 defects on player 2, player 3 may learn about that event from player 2, and consequently player 3 may also defect on player 1 afterwards, strengthening the initial incentive not to defect. They give an algorithm for identifying the (unique) maximal set of players that can sustain cooperation. In experiments, they identify a phase transition where, as parameters gradually change, the cooperative set suddenly changes from all players to no players. This, then, seems a rather brittle solution to the problem of cooperation to hang our hat on.

[3] This equilibrium is also subgame perfect. Another well-known strategy for the repeated Prisoner's Dilemma is *tit for tat*, which cooperates in the first round and then plays whatever the other player played in the previous round. Tit for tat is famous for its success in tournaments run by Axelrod (1984). The equilibrium in which both players play tit for tat is not subgame-perfect, however.

Disarmament

In some cases, the game will be played only once and the stakes will be high, but we can prepare for it, by taking actions ahead of time that will change the strategic structure of the game for the better. A natural example of this is that the players can reduce the set of actions available to them. But generally it does not hurt a player to have more options available, so why would a player reduce her set of options? One reason is that the players alternatingly and verifiably reduce their options, and each does so for the reason that this will induce the other to continue with this process as well. This is the idea of *disarmament* in game theory (Deng and Conitzer 2017, 2018). Indeed, real-world disarmament – say, of nuclear warheads – is instructive to consider. Both sides prefer a situation in which both sides have fewer warheads, but getting there can be challenging: each side will want to make sure that the other side is in fact disarming as well, and does not want to end up in a situation where it has removed most of its warheads before the other side has done anything. Similarly (or really, more generally), when considering players disarming themselves in a game by sequentially reducing their strategy space, it is important that the one player has verifiably taken its intended strategy reduction step before the other player proceeds; and that these steps are scheduled carefully to give no player an incentive to halt disarmament at any point. Successful disarmament (in the sense of reaching a desirable outcome in equilibrium) cannot always be achieved, and it is in general NP-hard to determine whether it can. However, if it is possible to eliminate *mixed* strategies, a type of folk theorem holds that cooperation can always be attained (Deng and Conitzer 2017, 2018). The latter is arguably more natural in the context of AI, as lines of code can be alternately committed that prevent certain mixed strategies from being used.

Cooperation between copies

Imagine an AI system playing the Prisoner's Dilemma against a very similar opponent, e.g., an agent that was trained using the same learning algorithm on the same or similar data. Should the AI agent defect? Here is a line of reasoning that suggests cooperation: if the agent cooperates, chances are that the opponent agent will cooperate too; if the agent defects, chances are that the opponent will defect too; and since the agent favors mutual cooperation over mutual defection, the agent should cooperate (cf. Hofstadter 1983). Natural as this reasoning may seem, it is at odds with the standard analysis of the Prisoner's Dilemma, which says that you should defect because *no matter what the opponent does*, you are better off defecting.

Which line of reasoning one follows depends on what type of decision theory one endorses. *Evidential decision theory* favors cooperating, precisely because *conditional on cooperating, one expects to be better off* than conditional on defecting. In contrast, *causal decision theory* says that such reasoning is mistaken: while it may be true that the two agents' decisions are *correlated*, one agent's decision to cooperate is not *causing* the other to cooperate as well, and so the traditional game theoretic analysis remains correct.

Algorithm 1: Cooperative equilibrium by testing whether the opponent is a copy.

Input: This program p_i, opponent program p_{-i}
Output: Cooperate or Defect

```
if p_i = p_{-i} then
    return Cooperate
end if
return Defect
```

Algorithm 2: The ϵGroundedFairBot of Oesterheld (2019).

Input: This program p_i, opponent program p_{-i}
Output: Cooperate or Defect

```
With probability ε:
    return Cooperate
return p_{-i}(p_{-i}, p_i)
```

Ever since Nozick (1969) pointed out the conflict between CDT and EDT, the question of which of them is correct has been in contention (e.g., Gibbard and Harper 1976; Peterson 2009; Ahmed 2014; Oesterheld and Conitzer 2021); in fact, a number of theories other than CDT and EDT have in the meantime been proposed (e.g. Weirich 2016, Section 3.6; Levinstein and Soares 2020).

Of course, most *humans* never face such an extreme scenario. However, for software it is normal to be copied and so an AI agent may very well face a copy or a near-copy of itself (cf. Cavalcanti 2010, Sect. 5; Oesterheld 2021, Sect. 1; Conitzer 2019b). For example, different humans may instantiate different copies or near copies of the same software agent to act on their behalfs.

Unfortunately, the philosophical decision theory literature gives little guidance on how to design AI agents that implement, say, EDT. How can we build learning agents that, for instance, cooperate in a Prisoner's Dilemma against a copy? Some work already exists on this question. For example, Bell et al. (2021) show that softmax Q-learners behave more like CDT agents. Albert and Heiner (2001), Mayer, Feldmaier, and Shen (2016), and Oesterheld, Demski, and Conitzer (2021) describe methods of learning that result in EDT-like behavior (see also Oesterheld 2021). However, existing work generally has focused on toy scenarios like the Prisoner's Dilemma against an exact copy. We hope that future work in this area will shed light on the feasibility of building learning agents that cooperate against near-copies in complex asymmetric scenarios.

Cooperation by reading each other's code

Consider the strategic interactions between charitable funds. For example, imagine that Fund 1 and Fund 2 are both deciding how to allocate $10,000. (We may imagine that they are allocating their final $10,000 so that they cannot use repetition.) Both value the fight against global warming at a rate of 2 units of utility per dollar donated. However, Fund 1 values contributions to cleaning up the streets of Town 1 at 3 units per dollar, while Fund 2 values cleaning up the streets in Town 2 at 3 units per dollar. Then the two funds face a

Prisoner's Dilemma-like strategic problem. Specifically, in the unique equilibrium of the game, both funds clean up the streets in their local town. Meanwhile, they would both prefer it if they both gave to global warming charities, and would benefit from an arrangement to do so – e.g., if they could each make a donation to global warming charities that is conditioned on the other doing so as well (Conitzer and Sandholm 2011; Ghosh and Mahdian 2008; Kalai et al. 2010; Monderer and Tennenholtz 2009).

Now imagine that the charitable funds are managed by AI. Imagine further that the funds uses an *open-source* AI system, e.g., because (prospective) donors to the fund demand transparency or because the fund hopes that outsiders will help find software bugs in the AI's code. What does this mean for the strategic interaction described above? Can it help bring the funds to cooperation?

Playing games while being able to read one another's source code poses new theoretical challenges. (Some of the decision theory literature cited in the previous section is related, but gives little formal guidance.) For instance, one might think that an agent should try to predict the opponent's action and then play a best response. However, it is unclear how to predict the opponent when the opponent is also predicting you. For example, if each program first runs the opponent's program, the two programs fail to terminate. A second problem is that while in the Prisoner's Dilemma, playing a best response is possible without knowing anything about the opponent, doing so results in mutual defection. As we will see below, there's reason to hope for better.

To get a better grasp on open-source game theory, we now consider a particularly simple framework (Rubinstein 1998; Tennenholtz 2004). Imagine that Alice and Bob play a Prisoner's Dilemma, but instead of choosing actions (Cooperate or Defect), they each submit a computer program that in turn chooses an action. Importantly, the computer programs are given access to each other's source code before choosing an action. Alice and Bob then play a new normal-form game – which we call a *program game* – where each of them chooses from some set of computer programs and their utility is determined by the actions chosen by the two computer programs. The key idea is that we can thus avoid the perspective of the open-source agent itself, and instead adopt the external perspectives of Alice and Bob. In particular, we can analyze the game played by Alice and Bob with ordinary game-theoretic means such as Nash equilibrium.

It turns out that the program game has cooperative equilibria. The simplest one (given by Rubinstein (1998, Sect. 10.4) and Tennenholtz (2004), as well as earlier work by McAfee (1984) and Howard (1988)) has each player submit a program that cooperates if the opponent's program is equal to *this* program; see Algorithm 1. (Compare the above discussion of cooperation with copies.) Clearly, when both players submit this program, they both cooperate. If one of them, say Player 1, deviated to any other program, then Player 2 would defect. Hence, if both players submit the program of Algorithm 1, the players are in equilibrium.

Unfortunately, the practical relevance of this equilibrium is limited, because of how fragile it is. Both players have to submit the exact same program. If we consider the scenario of different AI-managed funds, the different AIs might necessarily have very different code. For instance, each fund's AI may have code specific to each of the fund's cause areas.

Recent work has tried to remedy this problem by coming up with more robust cooperative equilibria. Barasz et al. (2014) and Critch (2019) give equilibria in which the programs only cooperate if they can prove that their opponent also cooperates. A recent, even simpler proposal is ϵGroundedFairBot (Oesterheld 2019), a computer program for the Prisoner's Dilemma which cooperates with 1% probability and with the remaining 99% probability simulates the opponent and copies its action (Algorithm 2). When playing against itself, ϵGroundedFairBot terminates with probability 1 and cooperates. Moreover, if both players submit ϵGroundedFairBot, they are in equilibrium.

These results show that transparency allows for new cooperative equilibria. However, the program equilibrium perspective sidesteps the question of how the agent itself should learn or reason. Instead, it requires the original players (the designers) to be rational (cf. Demski and Garrabrant 2020, Sect. 2.2). Perhaps this is appropriate given that we in fact find ourselves in this perspective. However, a theory of how the agent itself should reason seems valuable not just from a theoretical but also from a practical perspective. Ideally, we could build a learning system that figures out on its own that it can use, say, ϵGroundedFairBot (compare the discussion of learning to play the Prisoner's Dilemma against similar opponents).

Self-locating beliefs

Consider again an AI agent that is reasoning strategically in an environment with other copies of itself. Another issue it may face is that *it does not even know which copy it is!* In fact, such uncertainty can be helpful in attaining cooperative behavior. Consider an exchange in which a variety of trading agents operate. Imagine that there is a particular trading strategy that is effective for the agent using it, but detrimental to the market as a whole, so we wish to outlaw it. Unfortunately, it is difficult to detect the strategy being used in the wild. However, a monitoring entity (say, the SEC) is able to place copies of the trading agents in simulated versions of the exchange, where the behavior is more easily detectable and the agent can be fined (in real dollars).[4] Thus, an agent is generally unsure whether it is in the real or the simulated environment, and may conclude that in expectation it is better not to use the strategy. How should it assess the probability that it is in the simulated environment? This question is studied in the literature on *self-locating beliefs*. We will now argue that this literature must play a key role in the game theory of highly advanced AI systems, as this issue is much more pervasive in this context than it may at first appear.

Questions of self-locating beliefs come up in scenarios where one or more agents are, or may be, in the same *epistemic state* across time and/or space – i.e., what they know, including about their own identity, is the same. (In game-theoretic parlance, these situations correspond to the game

[4] Cf. the use of honeypots in computer security (Spitzner 2002) and the Volkswagen emissions scandal.

tree having multiple nodes *in the same information set*, including along the same path down the tree.) There are multiple ways in which these issues come up for AI systems. First, an AI agent can be deliberately designed to forget its past, for example for privacy reasons, so that it repeatedly finds itself in the same epistemic state, confronted with the same decision to make without remembering that it has made that decision multiple times before. (Indeed, in the game theory literature, these types of issues are often considered under the heading of *imperfect recall*.) Similarly, the above example about inspecting a trading agent in a simulated environment could involve using the same instantiation of the agent, but first wiping its memory clean. Second, we can instantiate the same AI agent multiple times, by copying its code and running it in multiple places – and each instantiation may not know where or when it is being run. A special case of this, and one of particular interest to us, is that agent 1 is facing some (possibly entirely different) agent 2 in a strategic situation, has access to agent 2's source code, and decides to simulate agent 2. Then, reasoning from the perspective of agent 2, how can it be sure that it is not the simulated copy, and how should this affect agent 2's actions? We have already discussed the use of simulation by ϵGroundedFairBot to achieve cooperation above, but we did so mostly from an "external" perspective of choosing programs. Here, instead, we take the "internal" perspective of an agent that is potentially being simulated in such a scenario, because to fully analyze and understand these scenarios, we need to be able to reason from both these perspectives in a consistent manner.

The paradigmatic example in the literature on self-locating beliefs is the *Sleeping Beauty problem* (Elga 2000; Titelbaum 2013). Rather than present the traditional Sleeping Beauty case, we will discuss an AI-oriented version of it. Consider a car that is human-driven, but equipped with AI that detects serious problems (e.g., driver asleep at the wheel) and in such cases "wakes up" to take over control of the vehicle. Moreover, the vehicle does not retain any data about such events – perhaps intentionally, to avoid embarrassment or raised insurance rates. This makes it a game of imperfect recall, as the AI will not remember waking up before. Suppose that there are only two types of drivers in the world: half of them are good, and half of them are bad. With good drivers, there will be one serious problem during their ownership of the car; with bad drivers, two. Now imagine that the AI has just been woken up, and wishes to assess the odds that its driver is good (which may be relevant for determining when to pass control back to the driver). What should be the car's subjective belief as a probability (its *credence*) that the driver is good? As it turns out, the answer is not settled. Some people believe the credence should still be $1/2$; these people are called *halfers*. Others believe that having just been woken up provides some evidence that the driver is bad, and these people generally believe the answer is $1/3$; they are called *thirders*.

Similar cases can be given for copies and simulations of agents. Consider a standalone AI system that half of all households install on one device, and that the other households install on two devices; once installed, the AI assesses the probability that it is alone in the household. Or, consider a strategic setting with an agent that, with probability $1/2$, is not simulated by its opponent before it acts in the real world, and with the remaining probability, is simulated once by its opponent, to see what action it will take. When finding itself about to act, realizing it may just be being simulated by the opponent, the agent must assess the probability that the opponent never simulates it. (We will return to this scenario later in this section.) It can be argued that all three of these cases are equivalent to the Sleeping Beauty problem.

The Sleeping Beauty problem, and more broadly questions about self-locating beliefs, have received much attention in the philosophy literature due to their implications for a number of big questions. Might we ourselves be (in) a simulation? Specifically, if we believe that humans would eventually be able to, and choose to, simulate vast numbers of human lives, should we not think we ourselves are likely to be among the simulated? (This is called the "simulation argument" (Bostrom 2003).) Another question: do we have evidence that there will not be many trillions of humans in the universe, since if there were, one would be unlikely to find oneself among the first 100 billion or so to live? (This is called the "Doomsday argument."") These questions call for an understanding of how a given perspective in the world, such as one of a single human being at a specific point in time, is *sampled* from all such perspectives. Discussing such a sampling process may seem odd – one may hold that humans just exist across spacetime and that is that; nothing out there is "sampling" their experiences. Yet, e.g., Hellie (2013) has argued that such a neutral, "constellation" view of the universe does not match the experience that we are given; instead one finds oneself in an "embedded" (what we have called "internal") viewpoint and can reasonably ask, "why this one?"[5]

The question of how to assign self-locating beliefs has practical relevance. After all, AI agents need to make decisions in settings with uncertainty; and generally speaking, if one makes decisions based on wrongly calculated probabilities, those decisions are going to be worse. If these agents face problems of self-locating belief, we will want to know the correct answer to the Sleeping Beauty problem (and similar problems). Indeed perhaps decision scenarios with imperfect recall give us a way of resolving the Sleeping Beauty problem – surely not both $1/2$ and $1/3$ can yield correct decisions, and so we can see which one of them leads to worse decisions to rule that one out? (This requires adding a decision component to the Sleeping Beauty problem, for example having the agent bet on the number of awakenings.)

It turns out that it is not that simple, and in fact this leads us back to questions about decision theory. In particular, to a first approximation,[6] thirders will make good decisions if they use causal decision theory, and halfers will make good

[5]For more discussions on these types of questions, see, for example, Valberg (2007); Hare (2007, 2009, 2010); Merlo (2016); Conitzer (2019a, 2020).

[6]There are multiple variants of these decision theories that one can use, and it is important to use the right one (Conitzer 2015; Oesterheld and Conitzer 2022).

decisions if they use evidential decision theory (Hitchcock 2004; Draper and Pust 2008; Briggs 2010; Oesterheld and Conitzer 2022). The reason that the distinction between the decision theories is relevant here is that one's decision upon waking up is very good evidence for what one will do, or did do, on the other day in the case that there are two awakenings; after all, one is in the exact same epistemic state on both days.

We now return to AI agents that can reason about the probability that they are simulations, tying together a number of the themes so far. Consider the following game, known as the *trust game* or *investment game* in the literature (Berg, Dickhaut, and McCabe 1995): in stage 1, player 1 can choose to give some amount of money (say, \$10) to player 2, and if so this money is tripled before player 2 receives it (so player 2 receives \$30); then, in stage 2, player 2 can give any amount of money (say, \$15) back to player 1 (which this time is not tripled). A standard game-theoretic analysis of this game is that player 2 has no incentive to give back any money in the second stage, and player 1 can anticipate that and therefore has no incentive to give any money in the first stage. As in the Prisoner's Dilemma, this outcome is Pareto-dominated by other outcomes (such as the example outcome in parentheses above).[7] However, now imagine that the players are AI agents that can simulate each other. Suppose player 2 finds itself in the situation where it has just received \$30. Player 2 might then reason as follows: *Perhaps I am just a simulation, run by player 1 to determine whether I would in fact pay back anything. If that is so, I should pay back a good amount of money, so that player 1 will give to me in the real world, which is what I care about. Of course, I might also simply be player 2 in the real world, in which case I would rather pay back nothing.* To trade off between these scenarios, player 2 needs to reason about the *probability* of being a simulation, which, tying back to self-locating beliefs, perhaps depends on the number of times that player 1 simulates player 2 – which in turn requires a game-theoretic analysis to determine. How should we model these interactions, what are their equilibria like, and how should we compute these?

Equilibrium selection

Equilibrium *selection* remains an important and understudied topic, and one that cuts across all the preceding topics insofar as the techniques discussed there allow for multiple equilibria and do not make it clear which one is to be chosen. Most of the work in the AI literature so far focuses on being able to compute or learn *an* equilibrium, rather than, say, an optimal one.[8] There are various reasons for this. First, there are popular benchmarks, particularly two-player zero-sum games, in which any equilibrium strategy is as good as any other per the minimax theorem (von Neumann 1928). Second, computing just any equilibrium is of course easier than computing an optimal one. In the context of computing Nash equilibria of 2-player normal-form games, this is made precise by the fact that computing one Nash equilibrium is only PPAD-complete (Daskalakis, Goldberg, and Papadimitriou 2009; Chen, Deng, and Teng 2009) whereas computing an optimal Nash equilibrium is NP-hard (Gilboa and Zemel 1989; Conitzer and Sandholm 2008).

Perhaps more importantly, when multiple players each use certain *learning* algorithms, they are often guaranteed to converge to *an* equilibrium, but not an optimal one. In fact, experiments have shown that both in repeated games and in program games, the simpler Defect–Defect equilibrium is learned when using standard learning algorithms (e.g., Sandholm and Crites 1996; Foerster et al. 2018; Hutter 2020). A recent line of work in multi-agent reinforcement learning aims to develop algorithms that learn better equilibria (Foerster et al. 2018; Letcher et al. 2019).

In settings where we have no control over the agents and we are just trying to predict what might happen – which is often the viewpoint of economic theory – one generally would like to have a full picture of *all* equilibria, as opposed to just one or even just an optimal one. From the perspective of building advanced AI systems, however, we also need to be able to *coordinate* the players on a single (good) equilibrium to achieve good outcomes. Otherwise, even in a domain that allows good equilibria (e.g., repeated games), we run the risk that each player chooses to play according to an equilibrium that is, for example, just a bit better for that player than for the others, but the resulting profile of play may be neither an equilibrium nor a particularly good outcome. Hence, we need to address the equilibrium selection problem head-on rather than sidestep it.

Foundations of cooperative AI – a call to action

We hope to have made a convincing case for a research agenda on foundations of cooperative AI. This agenda is important due the risks and opportunities associated with increasing interactions among AI systems. While various existing individual research contributions are highly relevant to this agenda – including ones that we have not discussed here[9] – the agenda as a whole has received little study so far. This provides an opportunity to build on, bring together, and unify these various threads, as well as new ones. Finally, we hope to have made a good case that this direction is tractable.

References

Ahmed, A. 2014. *Evidence, Decision and Causality*. Cambridge University Press.

[7] The trust game is popular in the field of *behavioral game theory* (Camerer 2003), which studies how humans play games. Human players will generally cooperate (give in both roles) in experiments (Berg, Dickhaut, and McCabe 1995; Brülhart and Usunier 2012). Why do they cooperate? Perhaps they are worried about running into their partner again, or otherwise being judged on their behavior. However, they may also feel that not giving money back is simply *wrong*. This would be a normative reason to cooperate, and a corresponding game-theoretic solution concept is given by Letchford, Conitzer, and Jain (2008).

[8] There are, of course, exceptions, e.g., Sandholm, Gilpin, and Conitzer (2005), Zhang et al. (2022), and citations therein.

[9] One example is carefully designing the boundaries (Conitzer 2019b) or roles (Moon and Conitzer 2016) of agents to bring about cooperation.

Albert, M.; and Heiner, R. A. 2001. An Indirect-Evolution Approach to Newcomb's Problem. CSLE Discussion Paper, No. 2001-01.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety.

Andersen, G.; and Conitzer, V. 2013. Fast Equilibrium Computation for Infinitely Repeated Games. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 53–59. Bellevue, WA, USA.

Axelrod, R. 1984. *The Evolution of Cooperation*. Basic Books.

Barasz, M.; Christiano, P.; Fallenstein, B.; Herreshoff, M.; LaVictoire, P.; and Yudkowsky, E. 2014. Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic.

Basu, K. 1994. The Traveler's Dilemma: Paradoxes of Rationality in Game Theory. *American Economic Review*, 84(2): 391–395.

Bell, J.; Linsefors, L.; Oesterheld, C.; and Skalse, J. 2021. Reinforcement Learning in Newcomblike Environments. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 22146–22157. Curran Associates, Inc.

Berg, J.; Dickhaut, J.; and McCabe, K. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10: 122–142.

Borgs, C.; Chayes, J.; Immorlica, N.; Kalai, A. T.; Mirrokni, V.; and Papadimitriou, C. 2010. The myth of the Folk Theorem. *Games and Economic Behavior*, 70(1): 34–43.

Bostrom, N. 2003. Are We Living in a Computer Simulation? *The Philosophical Quarterly*, 53(211): 243–255.

Briggs, R. 2010. Putting a value on Beauty. In Tamar Szabó Gendler and John Hawthorne, ed., *Oxford Studies in Epistemology: Volume 3*, 3–34. Oxford University Press.

Brülhart, M.; and Usunier, J.-C. 2012. Does the trust game measure trust? *Economics Letters*, 115(1): 20–23.

Camerer, C. F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Cavalcanti, E. G. 2010. Causation, Decision Theory, and Bell's Theorem: A Quantum Analogue of the Newcomb Problem. *The British Journal for the Philosophy of Science*, 61(3): 569–597.

CFTC; and SEC. 2010. Findings Regarding the Market Events of May 6, 2010. Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues.

Chen, X.; Deng, X.; and Teng, S.-H. 2009. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM*, 56(3).

Conitzer, V. 2015. A Dutch book against sleeping beauties who are evidential decision theorists. *Synthese*, 192(9): 2887–2899.

Conitzer, V. 2019a. A Puzzle about Further Facts. *Erkenntnis*, 84(3): 727–739.

Conitzer, V. 2019b. Designing Preferences, Beliefs, and Identities for Artificial Intelligence. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 9755–9759. Honolulu, HI, USA.

Conitzer, V. 2020. The Personalized A-Theory of Time and Perspective. *Dialectica*, 74(1): 1–29.

Conitzer, V.; and Sandholm, T. 2008. New Complexity Results about Nash Equilibria. *Games and Economic Behavior*, 63(2): 621–641.

Conitzer, V.; and Sandholm, T. 2011. Expressive Markets for Donating to Charities. *Artificial Intelligence*, 175(7–8): 1251–1271.

Critch, A. 2019. 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.

Dafoe, A.; Bachrach, Y.; Hadfield, G.; Horvitz, E.; Larson, K.; and Graepel, T. 2021. Cooperative AI: machines must learn to find common ground. *Nature*, 593(7857): 33–36.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2020. Open Problems in Cooperative AI.

Daskalakis, C.; Goldberg, P.; and Papadimitriou, C. H. 2009. The Complexity of Computing a Nash Equilibrium. *SIAM Journal on Computing*, 39(1): 195–259.

Demski, A.; and Garrabrant, S. 2020. Embedded Agency.

Deng, Y.; and Conitzer, V. 2017. Disarmament Games. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 473–479. San Francisco, CA, USA.

Deng, Y.; and Conitzer, V. 2018. Disarmament Games with Resources. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, LA, USA.

Draper, K.; and Pust, J. 2008. Diachronic Dutch Books and Sleeping Beauty. *Synthese*, 164(2): 281–287.

Elga, A. 2000. Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60(2): 143–147.

Foerster, J.; Chen, R. Y.; Al-Shedivat, M.; Whiteson, S.; Abbeel, P.; and Mordatch, I. 2018. Learning with Opponent-Learning Awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 122–130.

Ghosh, A.; and Mahdian, M. 2008. Charity auctions on social networks. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1019–1028.

Gibbard, A.; and Harper, W. 1976. Counterfactuals and Two Kinds of Expected Utility. *Foundations and Applications of Decision Theory*, 1: 125–162.

Gilboa, I.; and Zemel, E. 1989. Nash and correlated equilibria: Some complexity considerations. *Games and Economic Behavior*, 1: 80–93.

Hare, C. 2007. Self-Bias, Time-Bias, and the Metaphysics of Self and Time. *The Journal of Philosophy*, 104(7): 350–373.

Hare, C. 2009. *On Myself, And Other, Less Important Subjects*. Princeton University Press.

Hare, C. 2010. Realism About Tense and Perspective. *Philosophy Compass*, 5(9): 760–769.

Hellie, B. 2013. Against Egalitarianism. *Analysis*, 73(2): 304–320.

Hitchcock, C. 2004. Beauty and the bets. *Synthese*, 139(3): 405–420.

Hofstadter, D. 1983. Dilemmas for Superrational Thinkers, Leading Up to a Luring Lottery. *Scientific American*, 248(6).

Howard, J. V. 1988. Cooperation in the Prisoner's Dilemma. *Theory and Decision*, 24: 203–213.

Hutter, A. 2020. Learning in two-player games between transparent opponents.

Kalai, A. T.; Kalai, E.; Lehrer, E.; and Samet, D. 2010. A commitment folk theorem. *Games and Economic Behavior*, 69(1): 127–137.

Kontogiannis, S. C.; and Spirakis, P. G. 2008. Equilibrium Points in Fear of Correlated Threats. In *Proceedings of the Fourth Workshop on Internet and Network Economics (WINE)*, 210–221. Shanghai, China.

Letcher, A.; Foerster, J.; Balduzzi, D.; Rocktäschel, T.; and Whiteson, S. 2019. Stable Opponent Shaping in Differentiable Games. In *ICLR 2019*.

Letchford, J.; Conitzer, V.; and Jain, K. 2008. An Ethical Game-Theoretic Solution Concept for Two-Player Perfect-Information Games. In *Proceedings of the Fourth Workshop on Internet and Network Economics (WINE)*, 696–707. Shanghai, China.

Levinstein, B. A.; and Soares, N. 2020. Cheating Death in Damascus. *The Journal of Philosophy*, 117(5).

Littman, M. L.; and Stone, P. 2005. A Polynomial-time Nash Equilibrium Algorithm for Repeated Games. *Decision Support Systems*, 39: 55–66.

Mayer, D.; Feldmaier, J.; and Shen, H. 2016. Reinforcement Learning in Conflicting Environments for Autonomous Vehicles. In *International Workshop on Robotics in the 21st Century: Challenges and Promises*.

McAfee, R. P. 1984. Effective Computability in Economic Decisions.

Merlo, G. 2016. Subjectivism and the Mental. *Dialectica*, 70(3): 311–342.

Monderer, D.; and Tennenholtz, M. 2009. Strong mediated equilibrium. *Artificial Intelligence*, 173(1): 180–195.

Moon, C.; and Conitzer, V. 2015. Maximal Cooperation in Repeated Games on Social Networks. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 216–223. Buenos Aires, Argentina.

Moon, C.; and Conitzer, V. 2016. Role Assignment for Game-Theoretic Cooperation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, 416–423. New York, NY, USA.

Nozick, R. 1969. Newcomb's Problem and Two Principles of Choice. In Nicholas Rescher et al., ed., *Essays in Honor of Carl G. Hempel*, 114–146. Springer.

Oesterheld, C. 2019. Robust Program Equilibrium. *Theory and Decision*, 86(1): 143–159.

Oesterheld, C. 2021. Approval-directed agency and the decision theory of Newcomb-like problems. *Synthese*, 198: 6491–6504.

Oesterheld, C.; and Conitzer, V. 2021. Extracting Money from Causal Decision Theorists. *Philosophical Quarterly*, 71(4). DOI 10.1093/pq/pqaa086.

Oesterheld, C.; and Conitzer, V. 2022. Can *de se* choice be *ex ante* reasonable in games of imperfect recall? Working paper.

Oesterheld, C.; Demski, A.; and Conitzer, V. 2021. A theory of bounded inductive rationality.

Peterson, M. 2009. *An Introduction to Decision Theory*. Cambridge University Press.

Rubinstein, A. 1998. *Modeling Bounded Rationality*. Zeuthen Lecture Book Series. The MIT Press.

Russell, S.; Vaidya, S.; and Bras, R. L. 2010. Machine Learning for Comprehensive Nuclear-Test-Ban Treaty Monitoring. *CTBTO Spectrum*, 14: 32–35.

Russell, S. J. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Sandholm, T.; Gilpin, A.; and Conitzer, V. 2005. Mixed-Integer Programming Methods for Finding Nash Equilibria. 495–501.

Sandholm, T. W.; and Crites, R. H. 1996. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37(1-2): 147–166.

Spitzner, L. 2002. *Honeypots: Tracking Hackers*. Addison Wesley.

Tennenholtz, M. 2004. Program equilibrium. *Games and Economic Behavior*, 49(2): 363–373.

Titelbaum, M. G. 2013. Ten reasons to care about the Sleeping Beauty problem. *Philosophy Compass*, 8(11): 1003–1017.

Valberg, J. J. 2007. *Dream, Death, and the Self*. Princeton University Press.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100: 295–320.

Weirich, P. 2016. Causal Decision Theory. In *The Stanford Encyclopedia of Philosophy*. Spring 2016 edition.

Zhang, B.; Farina, G.; Celli, A.; and Sandholm, T. 2022. Optimal Correlated Equilibria in General-Sum Extensive-Form Games: Fixed-Parameter Algorithms, Hardness, and Two-Sided Column-Generation. Available as arXiv:2203.07181.