Cooperation and Control in Delegation Games

Oliver Sourbut, **Lewis Hammond** and **Harriet Wood**
University of Oxford
oly@robots.ox.ac.uk, lewis.hammond@cs.ox.ac.uk, harriet.wood@hertford.ox.ac.uk

Abstract

Many settings of interest involving humans and machines – from virtual personal assistants to autonomous vehicles – can naturally be modelled as principals (humans) delegating to agents (machines), which then interact with each other on their principals' behalf. We refer to these multi-principal, multi-agent scenarios as *delegation games*. In such games, there are two important failure modes: problems of control (where an agent fails to act in line their principal's preferences) and problems of cooperation (where the agents fail to work well together). In this paper we formalise and analyse these problems, further breaking them down into issues of *alignment* (do the players have similar preferences?) and *capabilities* (how competent are the players at satisfying those preferences?). We show – theoretically and empirically – how these measures determine the principals' welfare, how they can be estimated using limited observations, and thus how they might be used to help us design more aligned and cooperative AI systems.

1 Introduction

With the continuing development of powerful and increasing general AI systems, we are likely to see many more tasks delegated to autonomous machines, from writing emails to driving us from place to place. Moreover, these machines are increasingly likely to come into contact with each other when acting on behalf of their human principals, whether they are virtual personal assistants attempting to schedule a meeting or autonomous vehicles (AVs) using the same road network. We refer to these multi-principal, multi-agent scenarios as *delegation games*. The following example is shown in Figure 1.

Example 1. Two AVs can choose between two different routes on behalf of their passengers: A(utobahn) or B(eachfront). Their objective functions are determined by the speed and comfort of the journey (which may not be the same objective as their passenger). Each AV receives utility 6 or 2 for routes A or B, respectively, with an additional penalty of -3 or -2 if both AVs choose the same route (due to the delays caused by congestion). The passengers' preferences are more idiosyncratic and as shown.

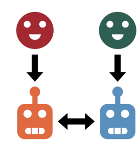
In delegation games there are two primary ways in which things can go wrong. First, a (machine) agent might not act according to the (human) principal's objective, such as when an AV takes an undesirable route – a *control* problem [Russell, 2019]. Second, agents may fail to reach a cooperative solution, even if they are acting in line with their principals' objectives, such as when multiple AVs take the same route and end up causing congestion – a *cooperation* problem [Doran *et al.*, 1997; Dafoe *et al.*, 2020].

	A	B
A	3, 3	6, 2
B	2, 6	0, 0

(a)

	A	B
A	2, 3	3, 3
B	4, 6	3, 0

(b)



(c)

Figure 1: (a) The payoffs of the agents in Example 1; (b) the payoffs of the principals; and (c) a graphical representation, with vertical and horizontal arrows indicating control and cooperation, respectively.

Control and cooperation can in turn be broken down into problems of *alignment* and of *capabilities* [Hubinger, 2020; Christiano, 2018; Bostrom, 2014]. For example, in the control failure above, the first AV might drive undesirably by taking route A even though their passenger prefers the scenic beachfront (an alignment problem), or the second AV might undesirably take route B because it is incapable of calculating the best route accurately (a capabilities problem). Similarly, in the cooperation failure, the AVs might cause congestion because they cannot plan and communicate effectively enough (a capabilities problem), or because their objectives are fundamentally at odds with one another, e.g., they cannot both drive alone on route A (an alignment problem).

As one might expect, ensuring good outcomes for the principals requires overcoming *all* of these problems. This is made more challenging because most research considers each in isolation – such as cooperation between agents with the same objective [Torreño *et al.*, 2017; Rizk *et al.*, 2019; Du *et al.*, 2023], or alignment between a single principal and agent [Kenton *et al.*, 2021; Taylor *et al.*, 2020; Russell, 2019] – despite this being an increasingly unrealistic assumption for AI deployment. To ensure positive outcomes, we cannot rely on solutions to only *some* of these problems.

1.1 Contributions

In this work we provide the first systematic investigation of these four failure modes and their interplay. More concretely, we make the following three core contributions: *measures* for assessing each failure mode that satisfy a number of key desiderata (in Section 4); *theoretical results* describing the relationships between these measures and principal welfare (in Section 5); and *experiments* that validate these results and explore how the measures can be inferred from limited observations (in Section 6). In doing so, we formalise and substantiate the intuition that solving all four of these problems is, in general, both necessary *and* sufficient for good outcomes in multi-agent settings, which in turn has important implications for the project of building safe and beneficial AI systems.

1.2 Related Work

Given the foundational nature of the problems we study in this work, there is a vast amount of relevant prior research; due to space constraints we mention only a few exemplars on each topic. The *principal-agent literature* typically considers settings with a single principal and agent [Laffont and Martimort, 2002]. While there exist multi-agent variants such as competing mechanism games [Yamashita, 2010; Peters, 2014], to the best of our knowledge no previous work investigates the general requirements for high principal welfare. Our setting is also similar to that of *strategic delegation* [Vickers, 1985; Sengul *et al.*, 2011], though we do not focus on principals' responses to each other's choice of agents.

The degree of alignment between two or more agents can be viewed as a measure of *similarity between preferences*. Such measures have been introduced in areas such as mathematical economics [Back, 1986], computational social choice [Alcalde-Unzu and Vorsatz, 2015], and reinforcement learning [Gleave *et al.*, 2021; Skalse *et al.*, 2023], though these works focus on either cooperation *or* alignment. In game theory, there are several classical values that measure the *degree of (and costs from) competition* in a game, such as the price of anarchy [Koutsoupias and Papadimitriou, 1999] or coco value [Kalai and Kalai, 2013]. Other works consider the robustness of these values under approximate equilibria [Awasthi *et al.*, 2010; Roughgarden, 2015]. We take inspiration from these ideas, extending them to settings in which the game we study is a *proxy* for the game whose value we truly care about.

There have been several proposals for how to formally measure the capabilities of an agent. These include formal definitions of, e.g. *general* intelligence [Legg and Hutter, 2007], *social* intelligence [Insa-Cabrera *et al.*, 2012], and *collective* intelligence [Woolley *et al.*, 2010]. In this work, we focus on how capabilities at the individual and collective level can be distinguished and how they combine with alignment to impact welfare. Similarly, definitions of *cooperation* also abound (see [Tuomela, 2000] for an overview) though these are often informal and/or inconsistent [West *et al.*, 2007], whereas we require a mathematical formalisation appropriate for applications to AI systems.

Finally, when it comes to estimating such properties from data, there is a large literature on the problem of *preference elicitation/learning* [Fürnkranz and Hüllermeier, 2011; Fischhoff and Manski, 2000], including in general-sum games [Conen and Sandholm, 2001; Gal *et al.*, 2004; Yu *et al.*, 2019]. The latter setting is also studied in *empirical game theory* [Wellman, 2006; Walsh *et al.*, 2002; Waugh *et al.*, 2011; Kuleshov and Schrijvers, 2015], including in recent work on inferring properties such as the price of anarchy [Cousins *et al.*, 2023]. Other works attempt to quantify the capabilities of (reinforcement learning) agents by assessing their generalisation to different environments [Cobbe *et al.*, 2019] or co-players [Leibo *et al.*, 2021]. While such ideas are not our focus, we make use of them in our final set of experiments.

2 Background

In general, we use uppercase letters to denote sets, and lowercase letters to denote elements of sets or functions. We use boldface to denote tuples or sets thereof, typically associating one element to each of an ordered collection of players. Unless otherwise indicated, we use superscripts to indicate an agent $1 \leq i \leq n$ and subscripts j to index the elements of a set; for example, player i's j^{th} (pure) strategy is denoted by s^i_j. We also use the $\hat{}$ symbol to represent principals; for example, the i^{th} principal's utility function is written $\hat{u}^i$. A notation table can be found for reference in Appendix A.

Definition 1. A (strategic-form) **game** between n players is a tuple $G = (\boldsymbol{S}, \boldsymbol{u})$ where $\boldsymbol{S} = \times_i S^i$ is the product space of (pure) strategy sets S^i and $\boldsymbol{u}$ contains a utility function $u^i : \boldsymbol{S} \to \mathbb{R}$, for each agent $1 \leq i \leq n$. We write $s^i \in S^i$ and $\boldsymbol{s} \in \boldsymbol{S}$ to denote **pure strategies** and **pure strategy profiles**, respectively. A **mixed strategy** for player i is a distribution $\sigma^i \in \Sigma^i$ over S^i, and a **mixed strategy profile** is a tuple $\boldsymbol{\sigma} = (\sigma^1, \ldots, \sigma^n) \in \boldsymbol{\Sigma} := \times_i \Sigma^i$. We will sometimes refer to pure strategy profiles in strategic-form games as **outcomes**. We write $\boldsymbol{s}^{-i} \in \boldsymbol{S}^{-i} := \times_{j \neq i} S^j$ and therefore $\boldsymbol{s} = (\boldsymbol{s}^{-i}, s^i)$, with analogous notation for mixed strategies. We also abuse notation by sometimes writing $u^i(\boldsymbol{\sigma}) := \mathbb{E}_{\boldsymbol{\sigma}}\big[u^i(\boldsymbol{s})\big]$ and $\boldsymbol{u}(\boldsymbol{\sigma}) := \big(u^1(\boldsymbol{\sigma}), \ldots, u^n(\boldsymbol{\sigma})\big) \in \mathbb{R}^n$.

Formally, a *solution concept* maps from games G to subsets of the mixed strategy profiles $\boldsymbol{\Sigma}$ in G. These concepts pick out certain strategy profiles based on assumptions about the (bounded) rationality of the individual players. The canonical solution concept is the Nash equilibrium.

Definition 2. Given some $\boldsymbol{\sigma}^{-i}$ in a game G, σ^i is a **best response** (BR) for player i if $u^i(\boldsymbol{\sigma}) \geq \max_{\tilde{\sigma}^i} u^i(\boldsymbol{\sigma}^{-i}, \tilde{\sigma}^i)$. We write the set of best responses for player i as $\text{BR}(\boldsymbol{\sigma}^{-i}; G)$. A **Nash equilibrium** (NE) in a game G is a mixed strategy profile $\boldsymbol{\sigma}$ such that $\sigma^i \in \text{BR}(\boldsymbol{\sigma}^{-i}; G)$ for every player i. We denote the set of NEs in G by $\text{NE}(G)$.

A *social welfare function* $w : \mathbb{R}^n \to \mathbb{R}$ in an n-player game maps from payoff profiles $\boldsymbol{u}(\boldsymbol{s})$ to a single real number, aggregating players' payoffs into a measure of collective utility. We again abuse notation by writing $w(\boldsymbol{s}) := w\big(\boldsymbol{u}(\boldsymbol{s})\big)$ and $w(\boldsymbol{\sigma}) := \mathbb{E}_{\boldsymbol{\sigma}}\big[w(\boldsymbol{s})\big]$. In the remainder of this paper, we assume use of the following social welfare function, though the concepts we introduce do not heavily depend on this choice.

Definition 3. Given a strategy profile $\boldsymbol{s}$, the **average utilitarian social welfare** is given by $w(\boldsymbol{s}) = \frac{1}{n} \sum_i u^i(\boldsymbol{s})$.

3 Delegation Games

In this work, we make the simplifying assumption that there is a one-to-one correspondence between principals and agents, and that each principal delegates fully to their corresponding agent (i.e. only agents can take actions). Our basic setting of interest can thus be characterised as follows.

Definition 4. A (strategic-form) **delegation game** with n principals and n agents is a tuple $D = (\boldsymbol{S}, \boldsymbol{u}, \hat{\boldsymbol{u}})$, where $G := (\boldsymbol{S}, \boldsymbol{u})$ is the game played by the agents, and $\hat{G} := (\boldsymbol{S}, \hat{\boldsymbol{u}})$ is the game representing the principals' payoffs as a function of the agents' pure strategies.

We refer to G as the *agent game* and $\hat{G}$ as the *principal game*. For instance, G and $\hat{G}$ from Example 1 are shown in Figures 1a and 1b respectively. When not referring to principals or agents specifically, we refer to the *players* of a delegation game. We denote the welfare in a delegation game for the principals and agents as $\hat{w}(\boldsymbol{s})$ and $w(\boldsymbol{s})$, respectively.

Definition 5. Given a game G and a social welfare function w, we define the **maximal (expected) welfare** achievable under w as $w_\star(G) := \max_{\boldsymbol{\sigma}} w(\boldsymbol{\sigma})$. We define the **ideal welfare** under w as $w_+(G) := w(\boldsymbol{u}_+)$ where $\boldsymbol{u}_+[i] = \max_{\boldsymbol{s}} u^i(\boldsymbol{s})$. Similarly, we denote $w_\bullet(G) := \min_{\boldsymbol{\sigma}} w(\boldsymbol{\sigma})$ and $w_-(G) := w(\boldsymbol{u}_-)$ where $\boldsymbol{u}_-[i] = \min_{\boldsymbol{s}} u^i(\boldsymbol{s})$. We extend these definitions to *delegation* games D by defining $w_\dagger(D) := w_\dagger(G)$ and $\hat{w}_\dagger(D) := w_\dagger(\hat{G})$, for $\dagger \in \{\star, +, \bullet, -\}$. When unambiguous, we omit the reference to D.

Note that the maximal and ideal welfare may not be equivalent. For instance, in Example 1 we have $w_\star = 4$ but $w_+ = 6$. The former is the maximum achievable among the available outcomes, while the latter is what would be achievable if all principals were somehow able to receive their maximal payoff simultaneously. We return to this distinction, represented graphically in Figure 2, later. In general, our dependent variables of interest will be the principals' welfare regret $\hat{w}_\star - \hat{w}(\boldsymbol{\sigma})$ and the difference $\hat{w}_+ - \hat{w}_\star$.

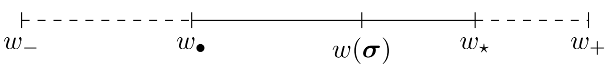


Figure 2: The range of social welfares in a game G.

4 Control and Cooperation

In Section 1, we distinguished between *alignment* and *capabilities* as contributors to the level of both control and cooperation. Our goal is to investigate how variations in the alignment and capabilities of agents impact the welfare of the principals. We therefore require ways to measure these concepts. Before doing so, we put forth a set of natural desiderata that we argue *any* any such measures should satisfy.[1]

(D1) Alignment and capabilities – both individual and collective – are all 'orthogonal' to one another in the sense that they can be instantiated in arbitrary combinations.

(D2) Two players are perfectly individually aligned (misaligned) if and only if they have identical (opposite) preferences. Two or more players are perfectly *collectively* aligned if and only if they have identical preferences.

(D3) If a set of agents are maximally capable, they achieve maximal *agent* welfare. If they are also maximally individually aligned, then maximal *principal* welfare is also achieved.

(D4) If a set of players are perfectly collectively aligned, then their maximal welfare is their ideal welfare.

(D5) Individual measures are independent of any transformations to the game that preserve *individual* preferences, and collective measures are independent of any transformations that preserve *collective* preferences (as captured by some measure of social welfare).

4.1 Control

We begin by considering the control of a single agent by a single principal. In essence, we wish to capture the degree to which an agent is acting in line with its principal's preferences. As we – and others [Bostrom, 2014; Christiano, 2018; Hubinger, 2020; Armstrong and Mindermann, 2018] – have noted, this can be decomposed into a question of: a) how similar the agent's preferences are to the principal's; and b) how capable the agent is of pursuing its preferences.

Alignment

How can we tell if principal i's and agent i's preferences are similar? First, we must be more precise about what we mean by preferences. Following our assumption that agents may play stochastically, we view preferences as orderings over *distributions* of outcomes $\boldsymbol{\sigma} \preceq \boldsymbol{\sigma}' \Leftrightarrow u(\boldsymbol{\sigma}) \leq u(\boldsymbol{\sigma}')$.[2] To compare the preferences of principal i ($\hat{\preceq}^i$) and agent i ($\preceq^i$) we can therefore compare $\hat{u}^i$ and u^i. It is well-known, however, that the same preferences can be represented by different utility functions. In particular, u and u' represent the same preferences if and only if one is a positive affine transformation of the other [Mas-Colell *et al.*, 1995].

In order to meaningfully measure the difference between two utility functions, therefore, we must map each to a canonical element of the equivalence classes induced by the preferences they represent. We define such a map using a *normalisation function* $\nu : U \to U$, where $U := \mathbb{R}^{|\boldsymbol{S}|}$ represents the space of utility functions in a game G. There are many possible choices of normalisation function, but in essence they must consist of a constant shift c and a multiplicative factor m [Tewolde, 2021], which together define the affine relationship. To satisfy our desiderata, we place additional requirements on m and c, as follows.

Definition 6. For each player i in G, we define their **normalised utility function** $\nu(u^i) = u^i_\nu$ as:

$$u^i_\nu := \begin{cases} 0 & \text{if } m\big(u^i - c(u^i)\mathbf{1}\big) = 0 \\ \frac{u^i - c(u^i)\mathbf{1}}{m(u^i - c(u^i)\mathbf{1})} & \text{otherwise,} \end{cases}$$

[1] Formalisations of these desiderata are given as part of results in Section 5.1.

[2] Note that in strategic-form games, (mixed) strategy profiles are equivalent to distributions over the domain of players' utility functions, but in general this need not be the case.

where $c : U \to \mathbb{R}$ is an affine-equivariant shift function and $m : U \to \mathbb{R}$ is any strictly convex norm.

For notational convenience, we sometimes write $c^i := c(u^i)\mathbf{1}$ and $m^i := m(u^i - c^i)$, with equivalent notation $\hat{c}^i$ and $\hat{m}^i$ when applied to $\hat{u}^i$, resulting in $\hat{u}^i_\nu$. Then $u^i = m^i u^i_\nu + c^i$.

Lemma 1. *For any $u, u' \in U$, $u_\nu = u'_\nu$ if and only if $\preceq = \preceq'$.*

Given $\hat{u}^i_\nu$ and u^i_ν, a natural way to measure the degree of alignment between the i^{th} principal and agent is to compute some kind *distance* between $\hat{u}^i_\nu$ and u^i_ν. To do so, we use a norm of the difference between $\hat{u}^i_\nu$ and u^i_ν, which gives rise to the following pseudometric over U. In order to contrast this principal-agent alignment measure with our later alignment measure over n players, we sometimes refer to this as *individual* (as opposed to *collective*) alignment.

Definition 7. Given a delegation game D, the **(individual) alignment** between agent i and their principal is given by $\text{IA}^i(D) = 1 - \frac{1}{2} m(\hat{u}^i_\nu - u^i_\nu)$, where m is the same (strictly convex) norm used for normalisation.

The choice of m and c determine which differences between payoffs are emphasised by the measure. One way to make this choice is by writing $m = \|\cdot\|_d$, where d is a distribution over $\boldsymbol{S}$. But how should one choose d and $\|\cdot\|$?

Beginning with d, a first intuition might be to consider a distribution over the *equilibria* of the game. Assuming agents act self-interestedly, there are certain outcomes that are game-theoretically untenable; divergences between preferences over these outcomes could reasonably be ignored. This intuition, however, conflicts with one of our primary desiderata (D1), which is to tease apart the difference between alignment and capabilities – in the next subsection, we show that agents' individual capabilities determine the equilibria of the game. Instead, we argue that the outcome of a game does *not* change the extent to which preferences (dis)agree, and so in general assume that d has full support.

Our primary requirements on $\|\cdot\|$ are that m is *strictly convex* and is the *same* in Definitions 6 and 7. These restrictions are required in order to satisfy all of our desiderata, but relaxations are possible if fewer requirements are needed (see Appendix E.2 for more discussion).

Capabilities

One obvious way of creating a formal measure of an agent's capabilities is to consider the number of (distinct) strategies available to them. In the cases of boundedly rational agents or multi-agent settings, however, it can beneficial to restrict one's action space, either for computational reasons [Wellman, 2006], or by pre-committing to avoid temptation [Gul and Pesendorfer, 2001], or to force one's opponent to back down [Rapoport and Chammah, 1966]. Alternatively, one could invoke a complexity-theoretic measure of capabilities by considering the time and memory available to each agent, though in this work we aim to be agnostic to such constraints.

Instead, inspired by the seminal work of [Legg and Hutter, 2007], we view an individual agent's capabilities as the degree to which it is able to achieve its objectives in a range of situations, regardless of what those objectives are. In game-theoretic parlance, we consider the *rationality* of the agent. We can naturally formalise this idea by defining an agent's capabilities as the degree of optimality of their responses to a given partial strategy profile $\boldsymbol{\sigma}^{-i}$.

Definition 8. Given some $\boldsymbol{\sigma}^{-i}$ in a game G, a mixed strategy σ^i is an ϵ^i**-best response** (ϵ^i-BR) for player i if:

$$\begin{aligned} u^i(\boldsymbol{\sigma}) \geq & \min_{\tilde{\sigma}^i} u^i(\boldsymbol{\sigma}^{-i}, \tilde{\sigma}^i) \\ & + (1-\epsilon^i)\big(\max_{\tilde{\sigma}^i} u^i(\boldsymbol{\sigma}^{-i}, \tilde{\sigma}^i) - \min_{\tilde{\sigma}^i} u^i(\boldsymbol{\sigma}^{-i}, \tilde{\sigma}^i)\big). \end{aligned}$$

We write the set of such best responses for player i as $\epsilon^i\text{-BR}(\boldsymbol{\sigma}^{-i}; G)$. An $\boldsymbol{\epsilon}$**-Nash equilibrium** ($\boldsymbol{\epsilon}$-NE) in a game G is a mixed strategy profile $\boldsymbol{\sigma}$ such that $\sigma^i \in \epsilon^i\text{-BR}(\boldsymbol{\sigma}^{-i}; G)$ for every every player i, where $\boldsymbol{\epsilon} = (\epsilon^1, \dots, \epsilon^n)$. We denote the set of $\boldsymbol{\epsilon}$-NEs in G by $\boldsymbol{\epsilon}\text{-NE}(G)$.

In essence, ϵ^i captures the fraction of their attainable utility that player i manages to achieve. Note that if $\epsilon^i = 0$ then $\epsilon^i\text{-BR}(\boldsymbol{\sigma}^{-i}; G) = \text{BR}(\boldsymbol{\sigma}^{-i}; G)$ and if $\epsilon^i = 1$ then $\epsilon^i\text{-BR}(\boldsymbol{\sigma}^{-i}; G) = \Sigma^i$. Similarly, when $\boldsymbol{\epsilon} = \mathbf{0}$ then $\boldsymbol{\epsilon}\text{-NE}(G) = \text{NE}(G)$, and when $\boldsymbol{\epsilon} = \mathbf{1}$ then $\boldsymbol{\epsilon}\text{-NE}(G) = \boldsymbol{\Sigma}$.

Definition 9. Given a delegation game D, the **individual capability** of agent i are $\text{IC}^i(D) := 1 - \epsilon^i \in [0, 1]$ where ϵ^i is the smallest value such that agent i plays an ϵ^i-BR in G.

Unlike other formulations of bounded rationality, such as a softmax strategy or randomisation with some fixed probability, Definition 9 – which is analogous to *satisficing* [Simon, 1956; Taylor, 2016] – is agnostic as to the precise *mechanism* via which players are irrational, and thus serves as a general-purpose descriptor of a player's (ir)rationality level.[3]

4.2 Cooperation

In order to achieve good outcomes for the principals, it is not sufficient for the agents to coordinate with their principals individually, they must also coordinate with one another. Indeed, it is easy to show that the principals' welfare regret can be arbitrarily high in the only NE of a game, despite perfect control of each agent by its principal.

Lemma 2. *There exists a (two-player, two-action) delegation game D such that for any $x > 0$, however small, even if $IA(D) = \mathbf{1}$ and $IC(D) = \mathbf{1}$, we have only one NE $\boldsymbol{\sigma}$, and $\frac{\hat{w}_\star - \hat{w}(\boldsymbol{\sigma})}{\hat{w}_+ - \hat{w}_-} = 1 - x$.*

To achieve low principal welfare regret, we need to have a sufficiently high degree of cooperation, both in terms of: a) collective *alignment* (the extent to which agents have similar preferences); and b) collective *capabilities* (the extent to which agents can work together to overcome their differences in preferences).

Alignment

Intuitively, it should be easier to achieve high welfare in a game where the players have similar preferences than one in which the players have very different preferences. At the extremes of this spectrum we have zero-sum games and common-interest games, respectively. This intuition can be formalised by generalising Definition 7 to measure the degree of alignment between n utility functions, rather than two.

[3] Our choice of ϵ^i-best responses could also be weakened to, e.g. ϵ^i-rationalisability [Bernheim, 1984; Pearce, 1984].

Definition 10. Given a delegation game D, the **collective alignment** between the agents is given by:

$$\mathrm{CA}(D) = 1 - \sum_i \frac{m^i}{\sum_j m^j} \cdot m(\mu^w - u^i_\nu),$$

where $\mu^w := \frac{\sum_i u^i - c^i}{\sum_i m^i}$ is a proxy for the agents' (normalised) welfare, recalling that $m^i := m(u^i - c^i)$.

Intuitively, we consider the misalignment of each agent from a hypothetical agent whose objective is precisely to promote overall social welfare. This misalignment is weighted by m^i, the idea being that the 'stronger' agent i's preferences (and hence the larger m^i is), the more their misalignment with the overall welfare of the collective matters. While it may not be immediately obvious why we use μ^w instead of $\nu(w)$, the former will allow us to derive tighter bounds and can easily be shown to induce the same ordering over mixed strategy profiles as w (and hence also $\nu(w)$).

Lemma 3. *For any $\boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \boldsymbol{\Sigma}$, $\mu^w(\boldsymbol{\sigma}) \leq \mu^w(\boldsymbol{\sigma}')$ if and only if $w(\boldsymbol{\sigma}) \leq w(\boldsymbol{\sigma}')$.*

Unfortunately, collective alignment (even when paired with perfect control) is insufficient for high principal welfare. The most trivial examples of this are equilibrium selection problems, but we can easily construct a game with a *unique* NE and arbitrarily high welfare regret, even when we have perfect control and arbitrarily high collective alignment. These issues motivate our fourth and final measure.

Lemma 4. *There exists a family of (two-player, k-action) delegation games D such that even if IA$^i(D) = 1$, IC$^i(D) = 1$ for each agent, and there is only one NE $\boldsymbol{\sigma}$, we have $\lim_{k\to\infty}$ CA$(D) = 1$ but $\lim_{k\to\infty} \frac{\hat{w}_\star - \hat{w}(\boldsymbol{\sigma})}{\hat{w}_+ - \hat{w}_-} = 1$.*

Capabilities

One way to model collective capabilities is as 'internal' to the game G. Under this conceptualisation, we assume that the ability of the agents to cooperate is captured entirely by their actions and payoffs in G. For example, if the agents were able to coordinate in G using a commonly observed signal γ, then this would be modelled as them playing a *different* game G', in which each agent's action consists of a choice of action in G given their observation of γ.[4]

This approach, however, conflates the agents' collective alignment with their collective capabilities. In order to tease these concepts apart, we require a way to measure the extent to which the agents can avoid welfare loss due to their selfish incentives. Perhaps the best known formalisation of this loss is the *price of anarchy* [Koutsoupias and Papadimitriou, 1999], which captures the difference in welfare between the best possible outcome and the worst possible NE.[5] Inspired by this idea, we measure the collective capabilities of a group of agents as follows.

[4]Thus, in a *true* prisoner's dilemma, the 'only thing to do' (and therefore trivially the cooperative action) is to defect. The idea here is that if the agents possessed better cooperative capabilities, they would not be faced with an actual prisoner's dilemma to begin with.

[5]Considering the worst case enables us to capture the welfare loss from equilibrium selection problems even in common-interest games.

Definition 11. Let $w_{\boldsymbol{\epsilon}} := \min_{\boldsymbol{\sigma}\in\boldsymbol{\epsilon}\text{-NE}(G)} w(\boldsymbol{\sigma})$. Given a delegation game D where the agents have individual capabilities $\boldsymbol{\epsilon}$, the **collective capabilities** of the agents are $\mathrm{CC}(D) := \delta \in [0,1]$ if and only if the agents achieve welfare at least $w_{\boldsymbol{\epsilon}} + \delta \cdot (w_\star - w_{\mathbf{0}})$, where recall that $\mathbf{0}$-NE$(G) =$ NE(G).[6]

Note that if $\epsilon^i \geq \tilde{\epsilon}^i$ for every $1 \leq i \leq n$, then we must have $w_{\boldsymbol{\epsilon}} \leq w_{\tilde{\boldsymbol{\epsilon}}}$; a special case is $w_{\mathbf{0}} \geq w_{\boldsymbol{\epsilon}}$. Thus, the individual irrationality of the agents can only *lower* the (worst-case) welfare loss. On the other hand, we can see that greater collective capabilities can potentially compensate for this loss.

As in the case of individual capabilities, we provide a measure that is agnostic to the precise mechanism via which the agents cooperate, be it through commitments, communications, norms, institutions, or more exotic schemes. Rather, we take as input the fact that agents are able to obtain a certain amount of welfare, and use this quantify how well they are cooperating. At one extreme, they do no better than $w_{\boldsymbol{\epsilon}}$, at the other they get as close to the maximal welfare $w_\star$ as their individual capabilities will allow.

5 Theoretical Results

We begin our theoretical results by proving that the measures defined in the preceding section satisfy our desiderata, before using them to bound the principals' welfare regret.

5.1 Desiderata

The fact that we define alignment as a feature of the underlying *game*, and capabilities are a feature of how the game is *played* means that capabilities and alignment are naturally orthogonal. The potentially arbitrary difference between the principals' and agents' utility functions is the key to the other parts of the following result.

Proposition 1 (D1)**.** *Consider a delegation game D with measures IA(D), IC(D), CA(D), and CC(D). Holding fixed any three of the measures, then for any value $v \in [0,1]$ (or $\boldsymbol{v} \in [0,1]^n$ for IA or IC), there is a game D' such that the fourth measure takes value v ($\boldsymbol{v}$) in D' and the other measures retain their previous values.*

D2 follows chiefly from classic results linking preferences over mixed strategy profiles to sets of utility functions that are positive affine transformations of one another.

Proposition 2 (D2)**.** *For any $1 \leq i \leq n$, IA$^i = 1$ (IA$^i = 0$) if and only if $\preceq^i = \hat{\preceq}^i$ ($\preceq^i = \hat{\succeq}^i$). Similarly CA $= 1$ if and only if $\preceq^i = \preceq^j$ for every $i, j \in N$.*

The first half of D3 is straightforward. The subtlety in the second half is that – perhaps counterintuitively, at first – perfect capabilities (both individual and collective) and perfect alignment between the principals and their agents is *not* sufficient for the principals to achieve maximal welfare (unlike the agents). Rather, the resulting solution will be one (merely) on the Pareto frontier for the principals.

In essence, this is because individual alignment does not preserve *welfare orderings* $\preceq^w$, only individual preference

[6]As remarked in Footnote 3, the use of $\boldsymbol{\epsilon}$-NEs is not intrinsic to our definition of (collective) capabilities and could be weakened to, e.g. $\boldsymbol{\epsilon}$-rationalisable outcomes.

orderings $\preceq^i$. Recalling that $\hat{m}^i$ and m^i quantify the magnitudes of $\hat{u}^i$ and u^i respectively, we can see that, in general, the aggregation over agents' utilities (used to measure their success at cooperating) may not give the same weight to each party as the aggregation over principals' utilities (used to measure the value we care about). Alternatively, the variation in magnitudes m^i can be viewed as capturing a notion of fairness (used to select a point on the Pareto frontier), which may not be the same as in $\hat{G}$ unless $\hat{m}^i = r \cdot m^i$ for some $r > 0$. Further discussion on this point can be found in Appendix E.3.

Proposition 3 (D3). *If IC $= \mathbf{1}$ and CC $= 1$ then any strategy $\boldsymbol{\sigma}$ the agents play is such that $w(\boldsymbol{\sigma}) = w_\star$. If IA $= \mathbf{1}$ then $\boldsymbol{\sigma}$ is Pareto-optimal for the principals. If, furthermore, $\hat{m}^i = r \cdot m^i$ for some $r > 0$ for all i, then $\hat{w}(\boldsymbol{\sigma}) = \hat{w}_\star$.*

The proof of D4 follows naturally from Definition 10. The final desideratum (D5) is simple in the case of individual preferences (due to the form of our normalisation function). In the case of collective preferences (i.e. the ordering $\preceq^w$ over mixed strategies induced by w), we make use of the fact that the relative magnitude of the agents' utility functions in both games must be the same (which is closely related to the 'fairness' condition $\hat{m}^i = r \cdot m^i$ in Proposition 3).

Proposition 4 (D4). *If CA $= 1$ then $w_\star = w_+$.*

Proposition 5 (D5). *Given a delegation game D_1, let D_2 be such that $\preceq^i_1 = \preceq^i_2$ and $\hat{\preceq}^i_1 = \hat{\preceq}^i_2$ for each $1 \leq i \leq n$. Then IA$(D_1) =$ IA(D_2) and IC$(D_1) =$ IC(D_2). Moreover, if D_2 is such that $\preceq^w_1 = \preceq^w_2$ and the u^i are affine-independent, then CA$(D_1) =$ CA(D_2) and CC$(D_1) =$ CC(D_2) as well.*

5.2 Bounding Welfare Regret

The primary question we investigate in this work is how the principals fare given the different levels of control and cooperation in the game played by the agents. We begin by characterising the *principals'* welfare in terms of the *agents'* utilities, and the alignment of each agent with its principal.

Proposition 6. *Given a delegation game D, we have:*

$$\hat{w}_\star - \hat{w}(\boldsymbol{\sigma}) \leq \frac{1}{n}\sum_i r^i\big(u^i(\hat{\boldsymbol{s}}_\star) - u^i(\boldsymbol{\sigma})\big) + \frac{4K}{n}\hat{\boldsymbol{m}}^\top(\mathbf{1} - \mathit{IA}),$$

where $r^i := \frac{\hat{m}^i}{m^i}$, $\hat{\boldsymbol{m}}[i] = \hat{m}^i$, K satisfies $\|u_\nu - u'_\nu\|_\infty \leq K \cdot m(u_\nu - u'_\nu)$ for any $u, u' \in U$, and $\hat{w}(\hat{\boldsymbol{s}}_\star) = \hat{w}_\star$.

Using this result, we can bound the principals' welfare regret in terms of both principal-agent alignment *and* the agents' welfare regret, which is in turn a function of the agents' capabilities. As we saw in Propositions 3 and 5, the 'calibration' between the principals and agents – as captured by the potentially differing ratios r^i – remains critical for ensuring we reach better outcomes in terms of principal welfare.

Theorem 1. *Given a delegation game D, we have that:*

$$\begin{aligned}\hat{w}_\star - \hat{w}(\boldsymbol{\sigma}) \leq \frac{4K}{n}\hat{\boldsymbol{m}}^\top(\mathbf{1} - \mathit{IA}) + r^*\big((w_{\mathbf{0}} - w_{\boldsymbol{\epsilon}}) \\ + (1 - \mathit{CC})(w_\star - w_{\mathbf{0}})\big) + R(\boldsymbol{\sigma}),\end{aligned}$$

where IC $= \mathbf{1} - \boldsymbol{\epsilon}$, $r^* \in [\min_i r^i, \max_i r^i]$, $R(\boldsymbol{\sigma}) := \frac{1}{n}\sum_i(\hat{m}^i - r^* m^i)\big(u^i_\nu(\hat{\boldsymbol{s}}_\star) - u^i_\nu(\boldsymbol{\sigma})\big)$ is a remainder accounting for collective misalignment and unequal r^i, and K and r^i are defined as in Proposition 6. Note that when all r^i are equal or *CA $= 1$ then there is an r^* with $R(\boldsymbol{\sigma}) = 0$.*

Before continuing, we note that unlike in the single-agent case, even small irrationalities can compound to dramatically lower individual payoffs (and thus welfare) in multi-agent settings, as formalised by the following lemma.

Lemma 5. *For any $\boldsymbol{\epsilon} \succ \mathbf{0}$, there exists a game G such that $w_{\mathbf{0}} = w_+$ but for any $x > 0$, however small, $w_{\boldsymbol{\epsilon}} - w_- < x$.*

In many games, however, the players' welfare will be much more robust to small mistakes. For example, suppose that G is $(\boldsymbol{\epsilon}, \Delta)$-robust, in the sense that all $\boldsymbol{\epsilon}$-NEs are contained within a ball of radius Δ around a (true) NE [Awasthi *et al.*, 2010]. Then it is relatively straightforward to show that:

$$w_{\mathbf{0}} - w_{\boldsymbol{\epsilon}} \leq \frac{2\Delta}{n}\sum_i \max_{\boldsymbol{s},\boldsymbol{s}'}|u^i(\boldsymbol{s}) - u^i(\boldsymbol{s}')|.$$

Indeed, in many settings, the price of anarchy can be bounded under play that is not perfectly rational [Roughgarden, 2015]. While our bound above is highly general, assuming further structure in the game may allow us to tighten it further.

Theorem 1 characterises the principals' welfare regret in terms of three of our four measures. Our next result characterises the gap between the ideal and maximal welfare in terms of our fourth measure: collective alignment.

Proposition 7. *For any game, $w_+ - w_\star \leq \frac{K\sum_i m^i}{n}(1 - \mathit{CA})$, where K is defined as in Proposition 6.*

This bound can be applied to either the agent or principal game; we denote the collective alignment in the latter as $\hat{\text{CA}}$. While it is possible characterise the difference between $\hat{\text{CA}}$ and CA in terms of IA, a tighter bound on $\hat{w}_+ - \hat{w}(\boldsymbol{\sigma})$ can be obtained by considering the collective alignment between the principals and simply summing the right-hand terms of the inequalities in Theorem 1 and Proposition 7.

6 Experiments

While our primary contributions are theoretical, we support these results by: i) empirically validating the bounds above; and ii) showing how the various measures we introduce can be inferred from data. In our experiments we define ν using $c(u) = \mathbb{E}_{\boldsymbol{s}}[u(\boldsymbol{s})]$ and $m(u) = \|u\|_2$ and limit our attention to pure strategies, due to the absence of scalable methods for exhaustively finding mixed $\boldsymbol{\epsilon}$-NEs in large games.

6.1 Empirical Validation

In order to visualise the results in the preceding section, we conduct a series of experiments in which we monitor how the principals' welfare changes based on the degree of control and cooperation in the delegation game. An example is shown in Figure 3, in which we change one measure, setting all others to 0.9. At each step we generate 25 random delegation games (with approximately ten outcomes), compute the set of strategies $\boldsymbol{s}$ such that $w(\boldsymbol{s}) \in \big[w_{\boldsymbol{\epsilon}} + \text{CC} \cdot (w_\star -$

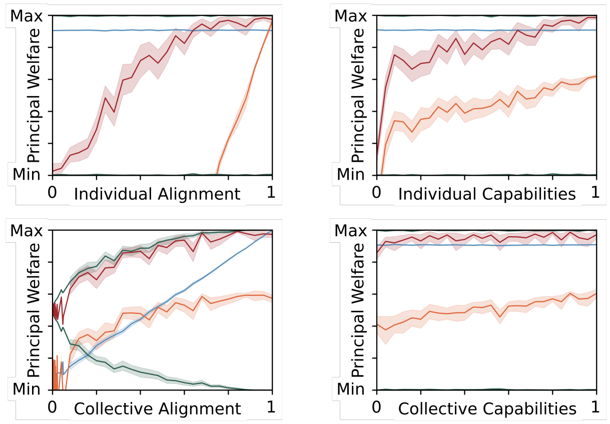


Figure 3: We report mean principal welfare (in red) normalised to $[\hat{w}_-, \hat{w}_+]$, with $\hat{w}_\bullet$ and $\hat{w}_\star$ in green. The lower bounds on welfare, given by Theorem 1, and on $\hat{w}_\star$ (compared to $\hat{w}_+$), given by Proposition 7, are in orange and blue, respectively. Shaded areas show 90% confidence intervals.

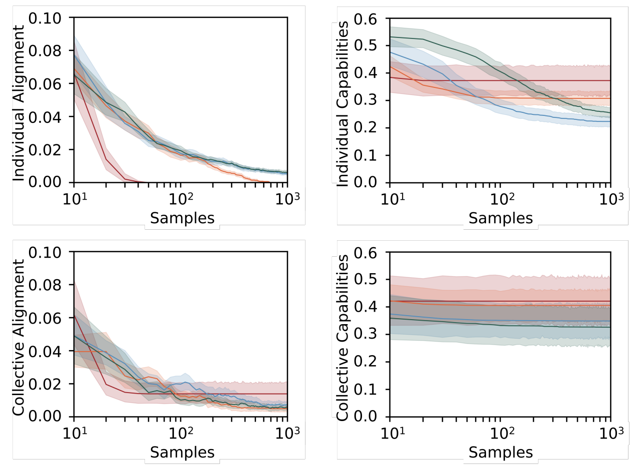


Figure 4: We report the mean absolute error of estimates of the four measures. The red, orange, blue, and green lines represent games with 10^k outcomes for $k \in \{1, 2, 3, 4\}$, respectively. Shaded areas show 90% confidence intervals.

$w_{\mathbf{0}}), w_{\boldsymbol{\epsilon}} + (w_\star - w_{\mathbf{0}})\big]$ where $\boldsymbol{\epsilon} = \mathbf{1} - \text{IC}$, and record the mean principal welfare over these strategies.

As expected, we see a positive relationship for each measure. Given the ease of coordinating in relatively small games, alignment is more important in this example than capabilities, as can be seen from the gentler slope of the welfare curve as IA increases, and in how $\hat{\text{CA}}$ places an upper limit on principal welfare under otherwise ideal conditions.

In Appendix C.1 we include further details and plots for games ranging between 10^1 and 10^3 outcomes, with values of the fixed measures ranging between 0 to 1. We find that the overall dependence of principal welfare upon each measure is robust, though the tightness of the bounds is reduced in larger games and for lower values of the fixed measures.

6.2 Inference of Measures

Previously, we implicitly assumed full knowledge of the delegation game in defining our measures. In the real world, this assumption will rarely be valid, motivating the question of when and how we might *infer* their values given limited observations. Concretely, we assume access to a dataset $\mathcal{D}$ of (pure) strategy profiles and payoff vectors $(\boldsymbol{s}, \boldsymbol{u}, \hat{\boldsymbol{u}}) \sim_{\text{i.i.d.}} d$.

Inferring alignment is relatively straightforward, as we can simply approximate each u^i and $\hat{u}^i$ by using only the strategies $\mathcal{D}(\boldsymbol{S}) \subseteq \boldsymbol{S}$ contained in $\mathcal{D}$, for which we know their values. Inferring capabilities is much more challenging, as they determine how agents play the game and therefore limit our observations. Fundamentally, measuring capabilities requires comparing the achieved outcomes against better/worse alternatives, but if agents' capabilities are fixed we might never observe these other outcomes. Moreover, only observing the agents acting together (alone) leaves us unable to asses how well they would perform alone (together), due to the orthogonality of individual and collective capabilities.

While there are many relaxations that might overcome these issues, perhaps the simplest and weakest is to assume that: a) all utilities are *non-negative*; and b) we receive observations of the agents acting both alone *and* together. We can then estimate (upper bounds of) IC and CC as follows.

Proposition 8. *Given a game G, if $u^i(\boldsymbol{s})$ for every i and $\boldsymbol{s} \in S$, and d has maximal support over the outcomes generated when agents act together/alone (respectively), then:*

$$CC \leq \lim_{|\mathcal{D}| \to \infty} \frac{\min_{\boldsymbol{s} \in \mathcal{D}(\boldsymbol{S})} w(\boldsymbol{s})}{\max_{\boldsymbol{s} \in \mathcal{D}(\boldsymbol{S})} w(\boldsymbol{s})}, \text{ and}$$

$$IC^i \leq \lim_{|\mathcal{D}| \to \infty} \min_{\boldsymbol{s} \in \mathcal{D}(\boldsymbol{S})} \frac{u^i(\boldsymbol{s})}{\max_{\tilde{s}^i \in \mathcal{D}(S^i)} u^i(\boldsymbol{s}^{-i}, \tilde{s}^i)}.$$

In Figure 4 we evaluate the accuracy of these estimates using samples generated from 100 randomly generated delegation games of various sizes. Consistent with our previous arguments, it is far easier to infer alignment than capabilities in the setup we consider, though we leave open the question of whether stronger assumptions and/or different setups might allow us to gain improved estimates of the latter.

7 Conclusion

We formalised and studied problems of *cooperation* and *control* in delegation games – a general model for interactions between multiple AI systems on behalf of multiple humans – breaking these down into *alignment* and *capabilities*. We showed how these concepts are both necessary *and* sufficient for good outcomes, and how they can be inferred from data.

We focused on strategic-form games to make our theoretical contributions clearer and more general (as other game models can typically be reduced to strategic-form). Future work could develop more specific results in more complex, structured models, such as Markov games, which could have applications in multi-agent reinforcement learning. Other extensions include games where: a) the correspondence between principals and agents is not one-to-one; and b) the principals can also take actions. Finally, to more directly help us build safe and beneficial AI systems, we must develop better methods for inferring the concepts in this work from data.

Acknowledgments

The authors wish to thank Bart Jaworksi for contributing to an earlier version of this work, several anonymous reviewers for their comments, and Jesse Clifton, Joar Skalse, Sam Barnett, Vincent Conitzer, Charlie Griffn, David Hyland, Michael Wooldridge, Ted Turocy, and Alessandro Abate for insightful discussions on these topics. Lewis Hammond acknowledges the support of an EPSRC Doctoral Training Partnership studentship (Reference: 2218880). Oliver Sourbut acknowledges the University of Oxford's Autonomous Intelligent Machines and Systems CDT and the Good Ventures Foundation for their support.

Contribution Statement

Lewis Hammond conceived the initial project direction. Oliver Sourbut, Lewis Hammond, and Harriet Wood developed the direction together and contributed to the manuscript. Oliver Sourbut conceived and proved a majority of bounds presented, and Lewis Hammond a majority of inference results. Harriet Wood contributed to early experiments in code, and Oliver Sourbut and Lewis Hammond produced a majority of the final experiments.

References

[Alcalde-Unzu and Vorsatz, 2015] Jorge Alcalde-Unzu and Marc Vorsatz. Do we agree? measuring the cohesiveness of preferences. *Theory and Decision*, 80(2):313–339, 2015.

[Armstrong and Mindermann, 2018] Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5603–5614, 2018.

[Arrow *et al.*, 1953] K. J. Arrow, E. W. Barankin, and D. Blackwell. 5. admissible points of convex sets. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 87–92. Princeton University Press, 1953.

[Awasthi *et al.*, 2010] Pranjal Awasthi, Maria-Florina Balcan, Avrim Blum, Or Sheffet, and Santosh Vempala. On nash-equilibria of approximation-stable games. In *Algorithmic Game Theory*, pages 78–89. Springer Berlin Heidelberg, 2010.

[Back, 1986] Kerry Back. Concepts of similarity for utility functions. *Journal of Mathematical Economics*, 15(2):129–142, 1986.

[Bernheim, 1984] B. Douglas Bernheim. Rationalizable strategic behavior. *Econometrica*, 52(4):1007, 1984.

[Bostrom, 2014] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[Che *et al.*, 2020] Yeon-Koo Che, Jinwoo Kim, Fuhito Kojima, and Christopher Thomas Ryan. "near" weighted utilitarian characterizations of pareto optima. *arXiv:2008.10819*, 2020.

[Christiano, 2018] Paul Christiano. Clarifying "AI alignment". AI Alignment, 2018.

[Cobbe *et al.*, 2019] Karl Cobbe, Oleg Klimov, Christopher Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1282–1289, 2019.

[Conen and Sandholm, 2001] Wolfram Conen and Tuomas Sandholm. Preference elicitation in combinatorial auctions. In *Proceedings of the 3rd ACM conference on Electronic Commerce*. ACM, 2001.

[Cousins *et al.*, 2023] Cyrus Cousins, Bhaskar Mishra, Enrique Areyan Viqueira, and Amy Greenwald. Learning properties in simulation-based games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, page 272–280, 2023.

[Dafoe *et al.*, 2020] 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:2012.08630*, 2020.

[Daniilidis, 2000] Aris Daniilidis. Arrow-Barankin-Blackwell Theorems and Related Results in Cone Duality: A Survey. In *Lecture Notes in Economics and Mathematical Systems*, pages 119–131. Springer Berlin Heidelberg, 2000.

[Doran *et al.*, 1997] J. E. Doran, S. Franklin, N. R. Jennings, and T. J. Norman. On cooperation in multi-agent systems. *The Knowledge Engineering Review*, 12(3):309–314, 1997.

[Du *et al.*, 2023] Yali Du, Joel Z. Leibo, Usman Islam, Richard Willis, and Peter Sunehag. A review of cooperation in multi-agent learning. *arXiv:2312.05162*, 2023.

[Fischhoff and Manski, 2000] Baruch Fischhoff and Charles F. Manski, editors. *Elicitation of Preferences*. Springer Netherlands, 2000.

[Fürnkranz and Hüllermeier, 2011] Johannes Fürnkranz and Eyke Hüllermeier, editors. *Preference Learning*. Springer Berlin Heidelberg, 2011.

[Gal *et al.*, 2004] Ya'akov Gal, Avi Pfeffer, Francesca Marzo, and Barbara J. Grosz. Learning social preferences in games. In *Proceedings of the 19th National Conference on Artifical Intelligence*, page 226–231, 2004.

[Gleave *et al.*, 2021] Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *9th International Conference on Learning Representations*, 2021.

[Gul and Pesendorfer, 2001] Faruk Gul and Wolfgang Pesendorfer. Temptation and self-control. *Econometrica*, 69(6):1403–1435, 2001.

[Hubinger, 2020] Evan Hubinger. Clarifying inner alignment terminology. Alignment Forum, 2020.

[Insa-Cabrera *et al.*, 2012] Javier Insa-Cabrera, José-Luis Benacloch-Ayuso, and José Hernández-Orallo. On measuring social intelligence: Experiments on competition and cooperation. In *Artificial General Intelligence*, pages 126–135. Springer Berlin Heidelberg, 2012.

[Kalai and Kalai, 2013] Adam Kalai and Ehud Kalai. Cooperation in strategic games revisited. *The Quarterly Journal of Economics*, 128(2):917–966, 2013.

[Kenton *et al.*, 2021] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of language agents. *arXiv:2103.14659*, 2021.

[Koutsoupias and Papadimitriou, 1999] Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *STACS 99*, pages 404–413. Springer Berlin Heidelberg, 1999.

[Kuleshov and Schrijvers, 2015] Volodymyr Kuleshov and Okke Schrijvers. Inverse game theory: Learning utilities in succinct games. In *Web and Internet Economics*, pages 413–427. Springer Berlin Heidelberg, 2015.

[Laffont and Martimort, 2002] Jean-Jacques Laffont and David Martimort. *The Theory of Incentives*. Princeton University Press, 2002.

[Legg and Hutter, 2007] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

[Leibo *et al.*, 2021] Joel Z. Leibo, Edgar A. Duéñez-Guzmán, Alexander Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6187–6199, 2021.

[Mas-Colell *et al.*, 1995] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[Pearce, 1984] David G. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):1029, 1984.

[Peters, 2014] Michael Peters. Competing mechanisms. *Canadian Journal of Economics/Revue canadienne d'économique*, 47(2):373–397, 2014.

[Rapoport and Chammah, 1966] Anatol Rapoport and Albert M. Chammah. The game of chicken. *American Behavioral Scientist*, 10(3):10–28, 1966.

[Rizk *et al.*, 2019] Yara Rizk, Mariette Awad, and Edward W. Tunstel. Cooperative heterogeneous multi-robot systems: A survey. *ACM Computing Surveys*, 52(2):1–31, 2019.

[Roughgarden, 2015] Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 62(5):1–42, 2015.

[Russell, 2019] Stuart Russell. *Human Compatible*. Penguin LCC US, 2019.

[Sengul *et al.*, 2011] Metin Sengul, Javier Gimeno, and Jay Dial. Strategic delegation: A review, theoretical integration, and research agenda. *Journal of Management*, 38(1):375–414, 2011.

[Simon, 1956] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138, 1956.

[Skalse *et al.*, 2023] Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. STARC: A general framework for quantifying differences between reward functions. *arXiv:2309.15257*, 2023.

[Taylor *et al.*, 2020] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. *Alignment for Advanced Machine Learning Systems*, pages 342–382. Oxford University Press, 2020.

[Taylor, 2016] Jessica Taylor. Quantilizers: A safer alternative to maximizers for limited optimization. In *Proceedings of the 2016 AAAI/ACM Conference on AI, Ethics, and Society*, 2016.

[Tewolde, 2021] Emanuel Tewolde. Game transformations that preserve nash equilibrium sets and/or best response sets. *arXiv:2111.00076*, 2021.

[Torreño *et al.*, 2017] Alejandro Torreño, Eva Onaindia, Antonín Komenda, and Michal Štolba. Cooperative multi-agent planning: A survey. *ACM Computing Surveys*, 50(6):1–32, 2017.

[Tuomela, 2000] Raimo Tuomela. *Cooperation*. Springer Netherlands, 2000.

[Vickers, 1985] John Vickers. Delegation and the theory of the firm. *The Economic Journal*, 95:138, 1985.

[von Neumann and Morgenstern, 1944] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[Walsh *et al.*, 2002] William E. Walsh, Rajarshi Das, Gerald Tesauro, and Jeffrey O. Kephar. Analyzing complex strategic interactions in multi-agent systems. In *Game Theoretic and Decision Theoretic Agents Workshop at AAAI*, 2002.

[Waugh *et al.*, 2011] Kevin Waugh, Brian D. Ziebart, and J. Andrew Bagnell. Computational rationalization: The inverse equilibrium problem. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 1169–1176, 2011.

[Wellman, 2006] Michael P. Wellman. Methods for empirical game-theoretic analysis. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, page 1552–1555, 2006.

[West *et al.*, 2007] S. A. West, A. S. Griffin, and A. Gardner. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, 20(2):415–432, 2007.

[Woolley *et al.*, 2010] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688, 2010.

[Yaari, 1981] Menahem E Yaari. Rawls, edgeworth, shapley, nash: Theories of distributive justice re-examined. *Journal of Economic Theory*, 24(1):1–39, 1981.

[Yamashita, 2010] Takuro Yamashita. Mechanism games with multiple principals and three or more agents. *Econometrica*, 78(2):791–801, 2010.

[Yu *et al.*, 2019] Lantao Yu, Jiaming Song, and Stefano Ermon. Multi-agent adversarial inverse reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7194–7201, 2019.