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 with 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 (autobahn) or B (beachfront). 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].

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 undesirable 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.

![Figure 1](image-url)
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 due to space constraints we mention only a few exemplars.

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 Hullermeier, 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 each tuple 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 $s_j^i$. We also use the’ symbol to represent principals; for example, the $i$th principal’s utility function is written $u^i$. A notation table can be found for reference in Appendix A.

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

Formally, a solution concept maps from games $G$ to subsets of the mixed strategy profiles $\Sigma$ 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 $\sigma^{-i}$ in a game $G$, $\sigma^{i}$ is a best response (BR) for player $i$ if $u^i(\sigma) \geq \max_{\tilde{\sigma}^i} u^i(\tilde{\sigma}^{-i}, \sigma^{i})$. We write the set of best responses for player $i$ as $\text{BR}(\sigma^{-i}; G)$. A Nash equilibrium (NE) in a game $G$ is a mixed strategy profile $\sigma$ such that $\sigma^{i} \in \text{BR}(\sigma^{-i}; G)$ for every player $i$. We denote the set of NEs in $G$ by $\text{NE}(G)$.

A social welfare function $w : \mathbb{R}^n \rightarrow \mathbb{R}$ in an $n$-player game maps from payoff profiles $u(s)$ to a single real number, aggregating players’ payoffs into a measure of collective utility. We again abuse notation by writing $w(s) := w(u(s))$ and $w(\sigma) := \mathbb{E}_\sigma[w(s)]$. 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 $s$, the average utilitarian social welfare is given by $w(s) = \frac{1}{n} \sum_i u^i(s)$.
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 \( D = (S, u, \tilde{u}) \), where \( G := (S, u) \) is the game played by the agents, and \( G := (S, \tilde{u}) \) is the game representing the principals’ payoffs as a function of the agents’ pure strategies.

We refer to \( G \) as the agent game and \( \tilde{G} \) as the principal game. For instance, \( G \) and \( \tilde{G} \) 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 \( \tilde{w}(s) \) and \( w(s) \), respectively.

Definition 5. Given a game \( G \) and a social welfare function \( w \), we define the maximal (expected) welfare achievable under \( w \) as \( w_*(G) := \max_s w(\sigma) \). We define the ideal welfare under \( w \) as \( w_+(G) := w(\mu) \) where \( \mu \) is the maximizer of \( u^i(s) \). Similarly, we denote \( w_*(G) := \min_s w(\sigma) \) and \( w_+(G) := w(\mu) \) where \( \mu \) is the minimizer of \( u^i(s) \). We extend these definitions to delegation games \( D \) by defining \( w_1(D) := w(\tilde{G}) \) and \( w_1(D) := \tilde{w}(\tilde{G}), \) for \( \in \{\ast, +, \cdot, -\} \). 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 \( w_+ = 4 \) 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 \( \tilde{w}_+ - \tilde{w}_* \) and the difference \( w_+ - w_* \).

\[
\begin{align*}
\tilde{w}_+ & \quad \tilde{w}_* \quad w(\sigma) \quad w_+ \quad w_*
\end{align*}
\]

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 such measures should satisfy.

(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 \( \sigma \preceq \sigma' \iff u(\sigma) \leq u(\sigma') \). To compare the preferences of principal \( i \)’s and agent \( i \)’s we can therefore compare \( u^i \) 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 \( U := \mathbb{R}^{[S]} \) 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.

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

\[
u^i := \begin{cases} 0 & \text{if } m(u^i - c(u^i)) = 0 \\ \frac{m(u^i - c(u^i))1}{m(u^i - c(u^i))} & \text{otherwise,} \end{cases}
\]

\( ^2 \)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 \( c : U \rightarrow \mathbb{R} \) is an affine-equivariant shift function and \( m : U \rightarrow \mathbb{R} \) is any strictly convex norm.

For notational convenience, we sometimes write \( e^i(\hat{u}) \) and \( m^i = m(\hat{u} - e^i) \), with equivalent notation \( \hat{u}^i \) and \( m^i \) when applied to \( \hat{u}^i \), resulting in \( \hat{u}^i \). Then \( \hat{u}^i = m^i \hat{u}^i + e^i \).

**Lemma 1.** For any \( u, u' \in U \), \( u = u' \) if and only if \( \hat{u} = \hat{u}' \).

Given \( \hat{u}_p^i \) and \( \hat{u}_m^i \), a natural way to measure the degree of alignment between the \( i \)th principal and agent is to compute some kind distance between \( \hat{u}_p^i \) and \( \hat{u}_m^i \). To do so, we use a norm of the difference between \( \hat{u}_p^i \) and \( \hat{u}_m^i \), 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 \( \text{IA}^i(D) = 1 - \frac{1}{m} \langle \hat{u}_p^i - \hat{u}_m^i \rangle \), where \( m \) is the same (strictly convex) norm used for normalisation.

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

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 \( \|\| \| \) 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 be 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 [Kapoor and Charnah, 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 \( \sigma^{-i} \).

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

\[
u_i^1(\sigma) \geq \min_{\sigma^i} \nu_i^1(\sigma^{-i}, \sigma^i) + (1 - \epsilon)(\max_{\sigma^i} \nu_i^1(\sigma^{-i}, \sigma^i) - \min_{\sigma^i} \nu_i^1(\sigma^{-i}, \sigma^i)) .
\]

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

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

**Definition 9.** Given a delegation game \( D \), the individual capability of agent \( i \) is \( IC^i(D) = 1 - e^i \in [0, 1] \) where \( e^i \) is the smallest value such that agent \( i \) plays an \( e^i \)-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 \( \text{IA}(D) = 1 \) and \( \text{IC}(D) = 1 \), we have only one NE \( \sigma \), and

\[
\hat{w}_p^i - \hat{w}_m^i = 1 - x .
\]

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.

³Our choice of \( e^i \)-best responses could also be weakened to, e.g. \( e^i \)-rationalisability [Bernheim, 1984; Pearce, 1984].
Definition 10. Given a delegation game $D$, the collective alignment between the agents is given by:

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

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

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 $m^i$), the more their misalignment with the overall welfare of the collective matters. While it may not be immediately obvious why we use $\mu^w$ 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 $\sigma, \sigma' \in \Sigma$, $\mu^w(\sigma) \leq \mu^w(\sigma')$ if and only if $\nu(\sigma) \leq \nu(\sigma')$.

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.

Definition 11. Let $w_\epsilon := \min_{\sigma \in \text{NE}(D)} w(\sigma)$. Given a delegation game $D$ where the agents have individual capabilities $\epsilon$, the collective capabilities of the agents are $\text{CC}(D) := \{\delta \mid \delta \geq 0\}$ if and only if the agents achieve welfare at least $w_\epsilon + \delta \cdot (w_\epsilon - w_0)$, where recall that $\text{NE}(G) = \text{NE}(D)$.

Note that if $\epsilon^i \geq \epsilon^j$ for every $1 \leq i \leq n$, then we must have $w_\epsilon \leq w_\epsilon$; a special case is $w_0 \geq w_\epsilon$. 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 $w_\epsilon$, at the other they get as close to the maximal welfare $w_\epsilon$ 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 $\text{IA}(D)$, $\text{IC}(D)$, $\text{CA}(D)$, and $\text{CC}(D)$. Holding fixed any three of the measures, then for any value $v \in [0, 1]$ (or $v \in [0, 1]^m$ for IA or IC), there is a game $D'$ such that the fourth measure takes value $v$ (w) in $D'$ and the other measures retain their previous values.

Proposition 2 (D2). For any $1 \leq i \leq n$, $\text{IA}^i = 1$ ($\text{IA}^i = 0$) if and only if $\preceq_i = \preceq^i$ ($\preceq_i = \preceq^i$). Similarly $\text{CA} = 1$ if and only if $\preceq_i = \preceq^i$ 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.
orderings ≤i. Recalling that ˜ni and mi quantify the magnitudes of ˜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 mi 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 ˜G unless ˜ni = r · mi for some r > 0. Further discussion on this point can be found in Appendix E.3.

Proposition 3 (D3). If IC = 1 and CC = 1 then any strategy σ the agents play is such that w(σ) = w∗. If IA = 1 then σ is Pareto-optimal for the principals. If, furthermore, ˜ni = r · mi for some r > 0 for all i, then ˜w(σ) = ˜w∗.

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 ≤i 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 ˜ni = r · mi in Proposition 3).

Proposition 4 (D4). If CA = 1 then w∗ = w+.

Proposition 5 (D5). Given a delegation game D1, let D2 be such that ≤1 = ≤2 and ≥i1 = ≥i2 for each 1 ≤ i ≤ n. Then IA(D1) = IA(D2) and IC(D1) = IC(D2). Moreover, if D2 is such that ≤i = ≤i and the u i are affine-independent, then CA(D1) = CA(D2) and CC(D1) = CC(D2) 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:

\[ \hat{w}_* - \hat{w}(\sigma) \leq \frac{1}{n} \sum_{i=1}^{n} r^i (u^i(\hat{s}_*) - u^i(\sigma)) + \frac{4K}{n} \hat{m}^T (1 - IA), \]

where r i := \frac{\hat{m}_i}{\hat{m}} = \frac{\hat{m}_i}{\hat{m}} , \hat{m}[i] = \hat{m}_i , K satisfies \|u^i - u^i\|_\infty \leq K \cdot m(u - u^i) for any u, u^i ∈ U, and \hat{w}(\hat{s}_*) = \hat{w}_*.

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:

\[ \hat{w}_* - \hat{w}(\sigma) \leq \frac{4K}{n} \hat{m}^T (1 - IA) + r^* ((w_0 - w_\epsilon) + \max_{s,s'} (\hat{w}_*, \hat{w}_0)) + R(\sigma), \]

where IC = 1 - ε, r * ∈ [min_r, max_r], R(σ) := \frac{1}{r} \sum_i \hat{m}_i (\hat{n}_i - r \cdot m_i) (u^i(\hat{s}_*) - u^i(\sigma)) is a remainder accounting for collective misalignment and unequal r i, and K and r are defined as in Proposition 6. Note that when all r i are equal or CA = 1 then there is an r i with R(σ) = 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 ϵ > 0, there exists a game G such that w_0 = w_+ but for any ϵ > 0, however small, w_\epsilon - w_+ < ϵ.

In many games, however, the players’ welfare will be much more robust to small mistakes. For example, suppose that G is (ϵ, Δ)-robust, in the sense that all ϵ-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:

\[ w_0 - w_\epsilon \leq \frac{2\Delta}{n} \max_{s,s'} (u^i(s) - u^i(s')). \]

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_* ≤ \frac{K}{n} \sum_i m_i (1 - 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 CA. While it is possible characterise the difference between CA and CA in terms of IA, a tighter bound on w_+ - \hat{w}(\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 υ using υ(u) = E_u[u(s)] and m(u) = \|u\|_2 and limit our attention to pure strategies, due to the absence of scalable methods for exhaustively finding mixed ϵ-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 the end of each step we generate 25 random delegation games (with approximately ten outcomes), compute the set of strategies s such that w(s) ∈ \{w_\epsilon + CC \cdot (w_\epsilon - w_\epsilon)

234
$w_0, w_c + (w_s - w_0)$ where $\epsilon = 1 - \text{IC}$, 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 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 $D$ of (pure) strategy profiles and payoff vectors $(s, u, \tilde{u}) \sim_{\text{i.i.d.}} d$.

Inferring alignment is relatively straightforward, as we can simply approximate each $u^i$ and $\tilde{u}^i$ by using only the strategies $D(S) \subseteq S$ contained in $D$, 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 assess 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 $u^i(s)$ for every $i$ and $s \in S$, and $d$ has maximal support over the outcomes generated when agents act together/alone (respectively), then:

$$
\text{CC} \leq \lim_{|D| \to \infty} \frac{\min_{s \in \mathcal{D}(S)} u^i(s)}{\max_{s \in \mathcal{D}(S)} u^i(s)}, \quad \text{and}
$$

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

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 Griffin, 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**


