Principal-Agent Boolean Games

David Hyland[1] , **Julian Gutierrez**[2] and **Michael Wooldridge**[1]

[1]University of Oxford, UK
[2]Monash University, Australia

{david.hyland, mjw}@cs.ox.ac.uk, julian.gutierrez@monash.edu

Abstract

We introduce and study a computational version of the principal-agent problem – a classic problem in Economics that arises when a principal desires to contract an agent to carry out some task, but has incomplete information about the agent or their subsequent actions. The key challenge in this setting is for the principal to design a contract for the agent such that the agent's preferences are then aligned with those of the principal. We study this problem using a variation of Boolean games, where multiple players each choose valuations for Boolean variables under their control, seeking the satisfaction of a personal goal, given as a Boolean logic formula. In our setting, the principal can only observe some subset of these variables, and the principal chooses a contract which rewards players on the basis of the assignments they make for the variables that are observable to the principal. The principal's challenge is to design a contract so that, firstly, the principal's goal is achieved in some or all Nash equilibrium choices, and secondly, that the principal is able to verify that their goal is satisfied. In this paper, we formally define this problem and completely characterise the computational complexity of the most relevant decision problems associated with it.

1 Introduction

The principal-agent problem is a classic problem in economics. It arises when a principal aims to sub-contract a task to an agent (or agents) while having only imperfect information about the agent(s). The two key classes of principal-agent problems are *adverse selection* and *moral hazard*. In a setting of adverse selection, the principal may have partial information about the agents' types, which may be relevant to their ability to carry out the delegated task. For example, when an employer hires a new employee, they have only partial information about the skills and aptitude of the employee. Their skills and aptitude will then play a part in the performance of the delegated task. In moral hazard settings, upon which we focus in the present paper, the principal is unable to observe the actions taken by agents subsequent to the contract. In particular, the principal may be unable to directly

verify that the terms of the contract have been respected. For example, suppose we want to hire a cleaner to ensure that a particular building is kept clean, although we are only able to inspect the building at infrequent intervals: if the cleaner desires to minimise effort, why would they then not simply clean up only when an inspection is imminent? The key challenge here is to design a contract for the agent that aligns the incentives of the agent with those of the manager, so that the manager can have confidence that the agent will be rationally motivated to expedite the task at hand, even if the manager is unable to directly verify that this is indeed the case.

Principal-agent problems of this type are extremely relevant in Artificial Intelligence as well as in Multi-Agent Systems: the venerable Contract Net protocol, for example, deals with precisely the situation where a principal delegates tasks to an agent [Smith, 1980], and although they were not considered in the original work, informational asymmetries between the principal and the agent may of course occur.

Our present work investigates a computational version of the principal-agent problem based on the Boolean games framework [Harrenstein *et al.*, 2001]. Boolean games are non-cooperative games in which players each control a finite set of Boolean variables, and the pure strategies available to a player correspond to the set of possible assignments of truth or falsity to those variables. Preferences are captured by assigning each player a propositional logic formula that the player desires to see satisfied. In the variant of Boolean games that we adapt in the present paper, secondary preferences are defined by costs associated with assignments of truth or falsity – a player first seeks to satisfy their goal, and secondarily seeks to minimise costs [Wooldridge *et al.*, 2013].

Classical work in the area of moral hazard in teams fails to take into account the potential lexicographic or qualitative nature of agent preferences. To our knowledge, our work is the first to study multi-agent moral hazard problems in a game where agents have lexicographic qualitative and quantitative preferences. In our model, a principal has a goal they desire to see accomplished, expressed, as with the agents' goals, as a propositional formula. The principal chooses a contract that rewards individual agents on the basis of their observable variables. The fundamental question we ask is whether it is possible for the principal to design a contract such that, if agents rationally make choices (i.e., make choices that form a game-theoretic equilibrium), then, firstly, the principal's goal

will be accomplished, and secondly, the principal can formally and automatically *verify* that the goal is accomplished, under the assumption that agents have acted rationally.

We derive complexity results for two classes of problems: Can the principal formally verify whether or not their objective, given as a propositional logic formula over the observable variables in the game, is satisfied on some/all observations and can the principal design a contract such that their objective is satisfied on some/all possible observations.

2 Related Work

Moral hazard. Investigations of the moral hazard problem were initially motivated by the desire to better understand the relationship between risk and incentives in insurance contracts [Pauly, 1968; Arrow, 1971]. Early models of moral hazard sought to study how risks can be shared optimally between principal and agent using contracts [Michael and William, 1976; Holmström, 1979; Shavell, 1979; Grossman and Hart, 1992].[1] Extensions of these models to settings with *multiple agents* and a single principal were also developed, in which the presence of competition between agents must be taken into account [Holmstrom, 1982; Itoh, 1991; Che and Yoo, 2001]. More recently, computational approaches to contract design have extended the original models in various ways and taken an algorithmic lens on computing optimal contracts [Dütting *et al.*, 2019; Dütting *et al.*, 2021; Alon *et al.*, 2021; Guruganesh *et al.*, 2021; Gan *et al.*, 2022; Dütting *et al.*, 2023]. Multi-agent contracts have also been studied from this perspective, which introduces the challenge of reasoning about combinatorially large action spaces [Babaioff *et al.*, 2006; Emek and Feldman, 2012; Dütting *et al.*, 2022; Castiglioni *et al.*, 2023]. Common among these studies is the assumption that agent preferences are quantitative, and partial observability is primarily modelled as a noisy signal of the agents' actions, which are often represented as a quantitative degree of effort. In contrast, our model allows agents to have qualitative goals specified as propositional formulae, which are prioritised over their quantitative objectives and constrains the agents' decision-making to a discrete action space. Furthermore, the principal's objective in our model is a propositional logic formula, as opposed to a function of the observation signal. An example of a setting where this may be more appropriate is a scenario in which a set of tasks are divided among a group of autonomous agents, each of whom are capable of reliably completing their assigned tasks. Each task is associated with a cost of completion, and the agents must decide which tasks to complete based on their objectives and the contract payments that are offered under the limited principal's observational abilities.

Mechanism design. There are computational approaches to mechanism design in which approximation algorithms [Nisan and Ronen, 1999], program synthesis [Narayanaswami *et al.*, 2022], and satisfiability checking [Maubert *et al.*, 2021; Mittelmann *et al.*, 2022], among others, have been proposed to automatically construct optimal mechanisms. These

---

[1]See [Georgiadis, 2023] for an overview of moral hazard studies.

works assume a stronger degree of control by the principal over the agents' interactions and their utilities.

Lexicographic qualitative and quantitative preferences. Finally, there have been several studies devoted to understanding solution concepts in games with lexicographic qualitative and quantitative preferences [Chatterjee *et al.*, 2005; Bloem *et al.*, 2009; Gutierrez *et al.*, 2017; Gutierrez *et al.*, 2021; Bulling and Goranko, 2022], but these do not consider problems of incentive design. Other previous work has studied the idea of a principal introducing a taxation scheme into games where players have a very similar preference structure to the one we consider [Wooldridge *et al.*, 2013; Harrenstein *et al.*, 2014; Harrenstein *et al.*, 2017; Levit *et al.*, 2019]. Additionally, [Gutierrez *et al.*, 2019] study a related problem called 'equilibrium design', but they consider players with only quantitative objectives. A key assumption in such previous works was that the principal can *fully observe* the outcome $\vec{v}$ selected by the agents. However, the principal does not have this luxury in many real-world situations, and it is this issue that motivates the study of the principal-agent problem. Other studies of incomplete information in Boolean games introduce partial observability at the level of the agents, as opposed to an external principal who seeks to influence their behaviour [Ågotnes *et al.*, 2013; Clercq *et al.*, 2014].

3 Preliminaries

Notation. Where S is a set, we denote the powerset of S by $\mathbf{2}^S$. We make use of propositional logic over Boolean variables throughout this study. In this context, we will let $\mathbb{B} = \{\top, \bot\}$ denote the set of Boolean literals. Where S is a set of Boolean variables, we let $\mathbb{B}^S$ denote the set of possible *valuations*, which are combinations of truth assignments to each variable in S. Where T is also a set of Boolean variables and $v \in \mathbb{B}^S$ is a valuation in S, we will let $v|_T$ denote the *restriction* of v to T, which is the Boolean valuation $v' \in \mathbb{B}^{S \cap T}$ that agrees with v on all variables in $S \cap T$.

3.1 Boolean Games

In this study, we will model a multi-agent moral hazard problem in the context of one-shot Boolean Games with costs, as presented in [Wooldridge *et al.*, 2013]. A *Boolean Game with costs* (henceforth "Boolean game" or simply "game") is a structure

$$B = (N, \Phi, (\Phi_i)_{i \in N}, (\gamma_i)_{i \in N}, (c_i)_{i \in N}),$$

where: $N = \{1, \ldots, n\}$ is a set of *agents*; $\Phi = \{p_1, p_2, \ldots, p_m\}$ is a finite set of $m \geq n$ *Boolean variables*; For each $i \in N$, the set Φ_i is the non-empty set of *Boolean variables uniquely controlled by agent* i, such that the collection $(\Phi_i)_{i \in N}$ forms a partition of Φ; For each $i \in N$, formula γ_i is the *goal* of agent i, which is represented as a propositional formula over Φ; For each $i \in N$, with $c_i : \mathbb{B}^\Phi \to \mathbb{Q}_+$ we represent the *cost function* of i, which assigns to each valuation $\vec{v}$ of the variables in Φ a non-negative cost $c_i(\vec{v})$, indicating the cost incurred by i under the valuation. We let c_i^* denote the maximal cost that i can incur under their c_i.

A *strategy* $v_i : \Phi_i \rightarrow \{\top, \bot\}$ for agent i is defined as an assignment of truth values to every variable in Φ_i. Here, we will use 1 and 0 interchangeably with $\top$ and $\bot$, respectively, in the context of strategies. For each agent $i \in N$, let V_i be their set of possible strategies. A *strategy profile* or *outcome* is then a tuple of strategies $\vec{v} = (v_1, \ldots, v_n)$, where agents select their strategies simultaneously in the absence of communication and knowledge of what strategies other agents selected. Additionally, given an agent $i \in N$, a strategy profile $\vec{v}$, and an individual strategy $v_i' \in \mathbb{B}^{\Phi_i}$ for i, we let $(\vec{v}_{-i}, v_i')$ be the strategy profile obtained by replacing v_i in $\vec{v}$ with v_i', i.e., the strategy profile $(v_1, \ldots, v_{i-1}, v_i', v_{i+1}, \ldots, v_n)$. Given a game B, we also define its *costless* version as the game $B^0 = (N, \Phi, (\Phi_i)_{i \in N}, (\gamma_i)_{i \in N}, (c_i^0)_{i \in N})$ which is as game B, except that $c_i^0(\vec{v}) = 0$ for all $\vec{v} \in \mathbb{B}^\Phi$ and $i \in N$.

For a strategy profile $\vec{v}$ and a propositional logic formula φ over the set of Boolean variables Φ, we write $\vec{v} \models \varphi$ to mean that $\vec{v}$ satisfies φ, where $\models$ denotes the propositional satisfaction relation. If it holds that $\vec{v} \models \gamma_i$ for some agent $i \in N$, we say that agent i's goal is satisfied under $\vec{v}$ and i is a *winner* under $\vec{v}$. Any agent whose goal is not satisfied under a strategy profile $\vec{v}$ will be referred to as a *loser* under $\vec{v}$. We write $W(\vec{v})$ and $L(\vec{v})$ to denote the sets of winners and losers under a given strategy profile $\vec{v}$, respectively.

3.2 Utilities, Preferences, and Nash Equilibrium

In order to model both qualitative *and* quantitative preferences, we consider a model where agent preferences are defined according to a lexicographic ordering on goal satisfaction and their incurred costs. Specifically, an agent i prefers any outcome in which their goal γ_i is satisfied over any outcome in which it is not, no matter what cost they incur. Secondarily, each agent prefers to minimise the cost that they incur. Formally, we define the *utility function* u_i for each agent $i \in N$ given a strategy profile $\vec{v}$ as follows:

$$u_i(\vec{v}) = \begin{cases} 1 + c_i^* - c_i(\vec{v}) & \text{if } \vec{v} \models \gamma_i \\ -c_i(\vec{v}) & \text{otherwise.} \end{cases}$$

Observe that if an agent i achieves their goal, then their utility will lie in the range $[1, c_i^* + 1]$; if their goal is not achieved, then their utility will lie in the range $[-c_i^*, 0]$. Preference relations $\succeq_i$ over outcomes for each $i \in N$ are defined in the obvious way: $\vec{v}_1 \succeq_i \vec{v}_2$ if and only if $u_i(\vec{v}_1) \geq u_i(\vec{v}_2)$, with indifference relations $\sim_i$ and strict preference relations $\succ_i$ defined in the usual way.

The primary solution concept we will work with is the pure strategy Nash equilibrium. Formally, a strategy profile $\vec{v}$ is a *Nash equilibrium* of a Boolean game B with cost function c if there is no agent $i \in N$ and strategy $v_i' \in V_i$ such that $(\vec{v}_{-i}, v_i') \succ_i \vec{v}$. If such a strategy exists for an agent $i \in N$, we say that i has a *beneficial deviation* from $\vec{v}$ to $(\vec{v}_{-i}, v_i')$. Where B is a game, we write $\text{NE}(B)$ to denote the set of Nash equilibria of B. Additionally, if φ is a Boolean logic formula over Φ, we let $\text{NE}_\varphi(B) = \{\vec{v} \in \text{NE}(B) \mid \vec{v} \models \varphi\}$ denote the set of Nash equilibria of B which satisfy formula φ.

4 The Principal-Agent Verification Problem

In this study, we aim to formulate and analyse the multi-agent moral hazard problem in the context of Boolean games with costs, where hidden actions are modelled by the principal only being able to observe the values of some subset $\mathcal{O} \subseteq \Phi$, known as the *observable set*. An *observation* by the principal is defined to be an assignment of truth values to all variables in its observable set $\mathcal{O}$, and is denoted by $\vec{o}$.[2] In general, a single observation $\vec{o}$ may correspond to more than one underlying strategy profile $\vec{v}$ if not all of the variables are fully observable. Thus, the observable set $\mathcal{O}$ naturally gives rise to the possibility for strategy profiles to be *indistinguishable* from one another. This can be expressed formally as an observational equivalence relation $=_\mathcal{O}$, defined over the set of observations $\mathbb{B}^\mathcal{O}$ such that for all strategy profiles $\vec{v}, \vec{v}' \in V$, we say that $\vec{v}$ is *indistinguishable* from $\vec{v}'$, written $\vec{v} =_\mathcal{O} \vec{v}'$, if it is the case that $\vec{v}|_\mathcal{O} = \vec{v}'|_\mathcal{O}$. If it is *not* the case that $\vec{v} =_\mathcal{O} \vec{v}'$, then we say that strategy profile $\vec{v}$ is *distinguishable* from strategy profile $\vec{v}$ and write $\vec{v} \neq_\mathcal{O} \vec{v}'$ in that case.

The problem faced by the principal is to design a contract so that they are able to ensure/verify that their objective has been accomplished on some or all of the Nash equilibria under the contract. To begin with, we will first analyse the simple case where the principal does not intervene, but simply asks whether they can *verify* whether or not some or all Nash equilibria consistent with any observation will satisfy their goal, given as a propositional logic formula φ. The value of this problem is that it provides a baseline against which the principal can assess the benefit of any potential contract that is designed. In order to formally state these problems, we first make precise the notion of consistency. Formally, we say that a strategy profile $\vec{v}$ is *consistent* with an observation $\vec{o}$ if it is the case that $\vec{v}|_\mathcal{O} = \vec{o}$. Then, we define the *consistent set* of an observation $\vec{o}$ in a given Boolean game B in the following way: $\text{CONS}(B, \vec{o}) = \{\vec{v} \in \text{NE}(B) \mid \vec{v} \text{ is consistent with } \vec{o}\}$. Finally, we define the set $\text{ENV}(B, \vec{o}, \varphi) = \text{NE}_\varphi(B) \cap \text{CONS}(B, \vec{o})$ to be the set of Nash equilibria in game B that satisfy φ and are consistent with $\vec{o}$. The first decision problem of interest is then stated as follows:

> E-NASH VERIFIABILITY:
> *Given*: Game B, observation $\vec{o} \in \mathbb{B}^\mathcal{O}$, formula φ.
> *Question*: Is it the case that $\text{ENV}(B, \vec{o}, \varphi) \neq \emptyset$?

Whilst an answer to the E-Nash Verifiability problem provides useful information to the principal about the *possibility* of their goal being achieved on some Nash equilibrium that is consistent with an observation, it does not make any guarantees as to whether or not the principal's goal has been achieved, given what they have observed. To obtain such a guarantee, we require the natural counterpart to this problem – the *A-Nash Verifiability* problem[3]:

> A-NASH VERIFIABILITY:
> *Given*: Game B, observation $\vec{o} \in \mathbb{B}^\mathcal{O}$, formula φ.
> *Question*: Is the case that $\text{CONS}(B, \vec{o}) \subseteq \text{NE}_\varphi(B)$?

If the answer to the A-Nash Verifiability problem is "yes", then the principal can be assured that any Nash equilibrium

---

[2] A special case of this is when we have $\mathcal{O} = \emptyset$, where the principal cannot observe the values of any of the Boolean variables in the game. In this case, we can say that the principal makes a *null observation*, which is written as $\vec{o} = ()$.

[3] Notation E- and A- are used to indicate existential and universal reasoning about the set of Nash equilibria in a game respectively.

	1	0
1	(2, 0)	(3, 0)
0	(1, 1)	(1, 1)

	1	0
1	(10, 1)	(10, 2)
0	(1, 2)	(1, 1)

Figure 1: Two cost matrices representing the agents' costs incurred under different strategy profiles in two different Boolean games. The game represented by the left figure is referred to as B_1 and discussed in Example 1. The game represented by the right figure is referred to as B_2 and discussed in Example 2. Player 1 (the row player) controls variable p_1 and player 2 (the column player) controls variable p_2.

consistent with their observation satisfies their goal. The following complexity results of the E- and A-Nash Verifiability problems can be readily derived from existing results relating to Boolean games [Wooldridge *et al.*, 2013]:

Proposition 1. E-NASH VERIFIABILITY *is* Σ_2^p-*complete, and* A-NASH VERIFIABILITY *is* Π_2^p-*complete.*

Example 1. *To illustrate the differences between the E-Nash and the A-Nash verifiability problems, consider the game B_1 consisting of two players with $\Phi_1 = \{p_1\}$, $\Phi_2 = \{p_2\}$, goals $\gamma_1 = \top$, $\gamma_2 = \neg p_1$, cost function as given by the table on the left in Figure 1, and observable set $\mathcal{O} = \{p_1\}$. Now suppose that the principal's objective is $\varphi = p_2$ and consider the E- and A-Nash verifiability problems for each observation $\vec{o} \in \mathbb{B}^{\mathcal{O}} = \{0, 1\}$. Firstly, we can identify the Nash equilibria in B_1 as being $(0, 0)$ and $(0, 1)$, with only the latter satisfying φ. Under the observation $\vec{o}_1 = (1)$, there is no Nash equilibrium in B_1 that is consistent with $\vec{o}_1$, so the answer to E-Nash verifiability is "no" and the answer to A-Nash verifiability is "yes" for $\vec{o}_1$. However, under the observation $\vec{o}_2 = (0)$, it is the case that $\mathrm{CONS}(B, \vec{o}) \cap \mathrm{NE}_\varphi(B) \neq \emptyset$ but not the case that $\mathrm{CONS}(B, \vec{o}) \subseteq \mathrm{NE}_\varphi(B)$. Thus, the answer to E-Nash verifiability for $\vec{o}_2$ is "yes" while it is "no" for A-Nash verifiability.*

5 Contract Design

The verification problem studied thus far allows the principal to obtain helpful information about whether or not they can be confident that their goal was satisfied under any possible observation that is consistent with at least one Nash equilibrium. However, in the classical formulations of the moral hazard problem, a principal is tasked with *designing* a contract to align the agents' incentives with their own. In this study, the principal's limitation lies not in an inability to accurately observe the outcome of its observables, but rather in their ability to only observe some subset of the actions chosen, as well as the agents' lack of willingness to compromise on satisfying their goal, no matter what the incentives are. For this, we first introduce the concept of contracts, which amount to an alteration of the costs incurred by the agents for taking various actions. This is similar to the notion of taxation schemes introduced in [Wooldridge *et al.*, 2013], but with the critical distinction that in our setting, the principal can only reward agents based on the outcomes of *observable* variables.

Formally, a contract $\kappa = (\kappa_1, \ldots, \kappa_n)$ is a tuple of functions $\kappa_i : \mathbb{B}^{\mathcal{O}} \to \mathbb{Q}_{\geq 0}$ for each agent $i \in N$ that map each possible observation by the principal to a non-negative rational number. Let $\mathcal{K}$ denote the set of all contracts for a given game and let κ_i^* denote the highest possible payment that the

principal offers to each agent $i \in N$ over all possible observations. Then, given a game B and an observation set $\mathcal{O}$, a contract κ gives rise to a new utility function u_i^κ for each agent $i \in N$ given a strategy profile $\vec{v}$:

$$u_i^\kappa(\vec{v}) = \begin{cases} 1 + c_i^* + \kappa_i(\vec{v}|_{\mathcal{O}}) - c_i(\vec{v}) & \text{if } \vec{v} \models \gamma_i \\ -\kappa_i^* + \kappa_i(\vec{v}|_{\mathcal{O}}) - c_i(\vec{v}) & \text{otherwise.} \end{cases}$$

Defining the utility of agents under a given contract in this way captures two desirable properties. Firstly, it preserves the lexicographic preferences of agents – each agent is guaranteed to achieve a strictly positive utility if their goal is satisfied, whereas they can achieve a utility of at most zero if it is not. Secondly, regardless of whether an agent's goal is satisfied, their utility is strictly increasing in the payment received from the principal and strictly decreasing in the costs they incur. Given a game B, an observable set $\mathcal{O}$ and a contract κ, we define the *Boolean game induced by κ* to be the game B^κ which is as game B, but where each agent i's utility is given by u_i^κ. Given a contract κ, an agent $i \in N$ and two strategy profiles $\vec{v}^1, \vec{v}^2$, we write $\vec{v}^1 \succeq_i^\kappa \vec{v}^2$ if and only if $u_i^\kappa(\vec{v}^1) \geq u_i^\kappa(\vec{v}^2)$ and define $\succ_i^\kappa$ in the obvious manner.

5.1 Inducing and Eliminating Equilibria

Due to the partial observability of the agents' actions and the fact that agents will always prefer to achieve their goal than not to do so, there are limits to the ability of a principal to either introduce or eliminate equilibria from a given game. As a special case, note that if all agents' actions are fully observable ($\mathcal{O} = \Phi$), then the contract problem can be characterised by the game's hard and soft equilibria – respectively those that can not be eliminated under any contract and those that can [Wooldridge *et al.*, 2013; Harrenstein *et al.*, 2014; Harrenstein *et al.*, 2017]. Thus, we will focus on the case where $\mathcal{O} \subset \Phi$. In this scenario, it is not in general possible to characterise the manipulability of a given game B by analysing the qualitative structure of the underlying costless Boolean game, as the principal is no longer able to eliminate the cost considerations for all agents by designing a contract so that each agent's quantitative payoff is constant regardless of their actions. Thus, the ability of the principal to induce or eliminate equilibria by means of contract design will be critically dependent on the initial cost functions of the agents.

Example 2. *As an example of the importance of initial costs in this model, consider a game B_2 consisting of two players with goals $\gamma_1 = \gamma_2 = \top$, cost function as given by the table on the right in Figure 1, and observable set $\mathcal{O} = \{p_1\}$. Now suppose again that the principal's objective is $\varphi = p_2$. With $\bar{p}_x$, $x \in \{1, 2\}$, we denote $\neg p_x$. Firstly, note that the only Nash equilibrium outcome of B_2 is $(0, 0)$. Now, because the principal cannot observe p_2 directly, the only way for them to ensure that p_2 is rationally chosen by player 2 is to offer a contract to player 1 to make the choice of setting $p_1 = 1$ more rational than setting it to 0. However, in order to do so, such a contract κ must satisfy $\kappa_1(p_1) > \kappa_1(\bar{p}_1) + 9$, under which the new unique Nash equilibrium becomes $(1, 1)$. This illustrates a scenario in which it is lucrative for an employee to control what is observable by the employer so as to benefit from indirect incentivisation.*

Before proceeding, we find it useful to define some terminology [Harrenstein *et al.*, 2014]. Firstly, we define the set $\text{INIT}(B) = \text{NE}(B^0)$ of *initial equilibria* to be the set of Nash equilibria in the costless game B^0. Formally, we will say that given a game B and an observable set $\mathcal{O}$, a strategy profile $\vec{v}$ is an *inducible equilibrium* if there exists a contract κ such that $\vec{v} \in \text{NE}(B^\kappa)$, and we let $\text{IND}(B, \mathcal{O})$ be the set of inducible equilibria of a game B with respect to an observable set $\mathcal{O}$. Additionally, we will say that a strategy profile $\vec{v}$ is a *hard equilibrium* (with respect to $\mathcal{O}$), written $\vec{v} \in \text{HARD}(B, \mathcal{O})$, if $\vec{v} \in \text{NE}(B)$ and there is no contract κ such that $\vec{v} \notin \text{NE}(B^\kappa)$. Complementarily, we say that a strategy profile $\vec{v}$ is a *soft equilibrium*, written $\vec{v} \in \text{SOFT}(B, \mathcal{O})$, if $\vec{v} \in \text{NE}(B) \setminus \text{HARD}(B, \mathcal{O})$. It is easy to verify that the previously defined sets are related in the following way:

Proposition 2. *For all Boolean games B and observable sets $\mathcal{O}$, it holds that $\text{HARD}(B, \mathcal{O}) \subseteq \text{IND}(B, \mathcal{O})$, $\text{SOFT}(B, \mathcal{O}) \subseteq \text{IND}(B, \mathcal{O})$, and $\text{NE}(B^\kappa) \subseteq \text{IND}(B, \mathcal{O}) \subseteq \text{INIT}(B)$ for all contracts κ.*

To reason about the principal's ability to design effective contracts, a method of analysing the inducibility and eliminability of sets of strategy profiles is needed. Following the approach of [Harrenstein *et al.*, 2017], we will introduce different types of potential deviations between strategy profiles. Suppose that $\vec{v}, \vec{v}' \in V$ are two distinct strategy profiles such that $\vec{v}' = (\vec{v}_{-i}, v_i')$ for some $i \in N$ and $v_i' \in V_i \setminus \{v_i\}$. Then we say that:

- $\vec{v}'$ is an *initial deviation* for i from $\vec{v}$, written $\vec{v} \rightharpoonup_i \vec{v}'$, if $i \in W(\vec{v}) \Rightarrow i \in W(\vec{v}')$.

- $\vec{v}'$ is an *inducible deviation* for i from $\vec{v}$, written $\vec{v} \rightarrow_i \vec{v}'$, if there is a contract κ such that $\vec{v}' \succ_i^\kappa \vec{v}$. Under such a contract, we say that κ *induces* a deviation from $\vec{v}$ to $\vec{v}'$.

- $\vec{v}'$ is a *soft deviation* for i from $\vec{v}$, written $\vec{v} \leftrightharpoons_i \vec{v}'$, if both $\vec{v} \rightarrow_i \vec{v}'$ and $\vec{v}' \rightarrow_i \vec{v}$.

- $\vec{v}'$ is a *hard deviation* for i from $\vec{v}$, written $\vec{v} \rightrightarrows_i \vec{v}'$, if $\vec{v} \rightarrow_i \vec{v}'$ but not $\vec{v}' \rightarrow_i \vec{v}$.

Note that just because a strategy profile is an initial deviation from another strategy profile, this does not necessarily imply that a contract exists which could *induce* that deviation, i.e., convert it into a beneficial deviation. Instead, we need the stronger notion of *inducible deviations*, which is, in turn, used to define soft and hard deviations. The following proposition characterises the relationship between initial and inducible deviations. The proofs of the following three Propositions proceed via a straightforward case analysis using the definitions of the previously defined concepts, so we omit them for the sake of space. For the backward implications, contracts can be designed to offer $c_i^* + 1$ to an agent i in order to make a particular observation more appealing than another observation, under which no reward is offered.

Proposition 3. *Let B be a game, $\mathcal{O} \subseteq \Phi$ an observable set, and $\vec{v}, \vec{v}' \in V$ be two distinct strategy profiles such that $\vec{v}' = (\vec{v}_{-i}, v_i')$ for some $i \in N$ and $v_i' \in V_i \setminus \{v_i\}$. Then, $\vec{v} \rightarrow_i \vec{v}'$ if and only if $\vec{v} \rightharpoonup_i \vec{v}'$ and one of the following conditions hold: 1) $\vec{v} \neq_{\mathcal{O}} \vec{v}'$, or 2) $\vec{v} =_{\mathcal{O}} \vec{v}'$ and $\vec{v}' \succ_i \vec{v}$.*

This result highlights the fact that the ability to design contracts to induce deviations between two strategy profiles is wholly determined by the initial cost and goal structure of the game in the case where the strategy profiles are indistinguishable. Next, we present a characterisation of the set of inducible equilibria of a game, which tells a principal when it is possible to design a contract κ such that a particular outcome becomes a Nash equilibrium under κ:

Proposition 4. *Let B be a game, $\mathcal{O} \subseteq \Phi$ an observable set, and $\vec{v}$ a strategy profile. Then, $\vec{v} \in \text{IND}(B, \mathcal{O})$ if and only if the following conditions hold: 1) $\vec{v} \in \text{INIT}(B)$; and 2) For all agents $i \in N$, and all choices $v_i' \in V_i$ such that $(\vec{v}_{-i}, v_i') =_{\mathcal{O}} \vec{v}$ and $i \in W(\vec{v}) \Leftrightarrow i \in W(\vec{v}_{-i}, v_i')$, we have that $c_i(\vec{v}_{-i}, v_i') \geq c_i(\vec{v})$.*

Now, we focus on characterising the sets of hard and soft equilibria, which allows a principal to check whether an undesirable initial equilibrium can be eliminated from a game.

Proposition 5. *Let B be a game, $\mathcal{O} \subseteq \Phi$ an observable set, and $\vec{v}$ a strategy profile. Then, $\vec{v} \in \text{SOFT}(B, \mathcal{O})$ if and only if the following conditions hold: 1) $\vec{v} \in \text{INIT}(B)$; and 2) There is agent $i \in N$ and a strategy $v_i' \in V_i \setminus \{v_i\}$ such that $(\vec{v}_{-i}, v_i') \neq_{\mathcal{O}} \vec{v}$ and $i \in W(\vec{v}_{-i}, v_i') \Leftrightarrow i \in W(\vec{v})$.*

Because $\text{HARD}(B) = \text{NE}(B) \setminus \text{SOFT}(B)$ by definition, a characterisation of the set of hard equilibria in a game follows immediately from Proposition 5:

Corollary 1. *Let B be a game, $\mathcal{O} \subseteq \Phi$ a non-empty observable set, and $\vec{v}$ a strategy profile. Then, $\vec{v} \in \text{HARD}(B, \mathcal{O})$ if and only if the following conditions hold: 1) $\vec{v} \in \text{NE}(B)$; and 2) For all agents $i \in N$ and all strategies $v_i' \in V_i \setminus \{v_i\}$ such that $(\vec{v}_{-i}, v_i') \neq_{\mathcal{O}} \vec{v}$, we have $(\vec{v}_{-i}, v_i') \rightrightarrows_i \vec{v}$.*

So far, our analysis has focused on whether or not a single strategy profile is an inducible, hard, or soft equilibrium. However, it is not sufficient in general to consider individual strategy profiles in this way, as there may be several strategy profiles that the principal would like to induce or eliminate. The more general problem faced by the principal is thus how it can determine whether a contract can be designed to jointly eliminate several equilibria in a game.

Therefore, we will extend the concept of eliminability to sets of strategy profiles. Given a set $X \subseteq V$ of strategy profiles, we say that X is *eliminable* if there is a contract κ such that $X \cap \text{NE}(B^\kappa) = \emptyset$. Any such contract κ is then said to *eliminate* X. To characterise the ability of the principal to eliminate sets of equilibria, we will utilise a graph-based approach to represent inducible deviations between strategy profiles. Given a Boolean game B and an observable set $\mathcal{O}$, we define the *potential deviation graph* of B with respect to $\mathcal{O}$ to be the directed graph $\mathcal{G}(B, \mathcal{O}) = (\mathcal{V}, E)$, where

- $\mathcal{V}$ is the set of vertices, where each vertex corresponds to a strategy profile $\vec{v} \in V$;
- $E = \{(\vec{v}, \vec{v}') \in \mathcal{V} \times \mathcal{V} \mid \vec{v} \rightarrow_i \vec{v}' \text{ for some } i \in N\}$ is the set of directed edges, where there is an edge from a vertex $\vec{v} \in \mathcal{V}$ to another vertex $\vec{v}' \in \mathcal{V}$ if and only if it is the case that $\vec{v} \rightarrow_i \vec{v}'$ for some $i \in N$.

The potential deviation graph of a game B wrt. an observable set $\mathcal{O}$ represents all possible deviations that could be induced by a contract κ. We say that a subgraph $D = (\mathcal{V}_D, E_D)$

of $\mathcal{G}(B, \mathcal{O})$ is the *deviation graph* induced by a contract κ iff it is the case that for every pair $\vec{v}, \vec{v}' \in \mathcal{V}_D$, we have $(\vec{v}, \vec{v}') \in E_D \Leftrightarrow \vec{v}' \succ_i^\kappa \vec{v}$. Intuitively, a contract κ induces a deviation graph D when only those inducible deviations that are actually made beneficial for some $i \in N$ under κ are included as edges in D, along with both of their endpoints. For the purposes of eliminating undesirable equilibria, we will also need to determine, given a deviation graph D of $\mathcal{G}(B, \mathcal{O})$, whether there exists a contract κ such that D is the deviation graph induced by κ. If such a contract exists, we will say that D is an *inducible deviation graph* of B wrt. $\mathcal{O}$.

We say that a tuple $P = (\vec{v}^1, \vec{v}^2, \dots, \vec{v}^k)$ is a *deviation path* if it holds that $\vec{v}^1 \to_{i^1} \vec{v}^2 \to_{i^2} \dots \to_{i^{k-1}} \vec{v}^k$, for some tuple of agents $(i^1, i^2, \dots, i^{k-1})$. Equivalently, a deviation path may be defined simply as any directed path within the potential deviation graph $\mathcal{G}(B, \mathcal{O})$. A deviation path $(\vec{v}^1, \vec{v}^2, \dots \vec{v}^k)$ is a *deviation cycle* if $k \geq 2$ and $\vec{v}^1 = \vec{v}^k$. Additionally, we will say that a deviation path $(\vec{v}^1, \vec{v}^2, \dots, \vec{v}^k)$ *involves agent i* if and only if we have $\vec{v}^j \to_i \vec{v}^{j+1}$ for some $j \in \{1, \dots, k-1\}$ and $i \in N$.

Although deviation paths are defined with respect to sequences of strategy profiles that are pairwise related by inducible deviations, it will be remembered that the influence of the principal is restricted to providing incentives only based on what they can observe, rather than the strategy profile that is actually chosen. In particular, given two strategy profiles $\vec{v}, \vec{v}' \in V$ such that $\vec{v} \neq_\mathcal{O} \vec{v}'$ and $\vec{v} \to_i \vec{v}'$ for some $i \in N$, the principal cannot design a contract that rewards i for choosing $\vec{v}'$ over $\vec{v}$ directly, but this must instead be done indirectly, by rewarding i sufficiently more under the *observation $\vec{v}'|_\mathcal{O}$* than the observation $\vec{v}|_\mathcal{O}$. From the point of view of the principal, then, it will sometimes be more useful to reason about *observed deviation paths*, which we define as a tuple $P_\mathcal{O} = (\vec{o}^1, \vec{o}^2, \dots, \vec{o}^k)$ such that there exists a set $\{\vec{v}^1, \vec{v}^2, \dots, \vec{v}^k\} \subseteq \mathcal{V}$ where for all $j \in \{1, \dots, k\}$, we have 1) $\vec{v}^j|_\mathcal{O} = \vec{o}^j$ and 2) if $j < k$, then $(\vec{v}^j, \vec{v}^{j+1'}) \in E_D$ for some $\vec{v}^{j+1'} =_\mathcal{O} \vec{v}j + 1$. In other words, an observed deviation path need not be an actual deviation path in $\mathcal{G}(B, \mathcal{O})$; it simply needs to look like one from the principal's perspective. Given this, a deviation path $P = (\vec{v}^1, \vec{v}^2, \dots, \vec{v}^k)$ is said to contain an *observed deviation cycle* if $\vec{v}^1 =_\mathcal{O} \vec{v}^j$, for some $j \in \{2, \dots, k\}$. Terminology regarding the involvement of agents in observed deviation paths/cycles is extended in the natural way – an observed deviation path $P_\mathcal{O} = (\vec{o}^1, \vec{o}^2, \dots, \vec{o}^k)$ *involves agent i* iff $\vec{v}^j \to_i \vec{v}^{j+1}$ for some $j \in \{1, \dots, k-1\}$ such that $\vec{v}^j|_\mathcal{O} = \vec{o}^j$ and $\vec{v}^{j+1}|_\mathcal{O} = \vec{o}^{j+1}$.

Example 3. *To illustrate some of the recently introduced concepts, consider the two-player game in Figure 2. The diagram at the bottom depicts a deviation path $P = ((1, 0, 0), (0, 1, 0), (1, 1, 0))$ consisting of three Nash equilibria of B. Note that P involves only one agent and is not a deviation cycle. Nevertheless, P contains an observed deviation cycle that involves only player 1, as corresponding observed deviation path is $P_\mathcal{O} = (\vec{o}_1, \vec{o}_0, \vec{o}_1)$. We will see shortly that this property is crucial in determining whether a set of strategy profiles can be eliminated or not.*

Now, we present a characterisation of the eliminability of sets of initial equilibria in a game:

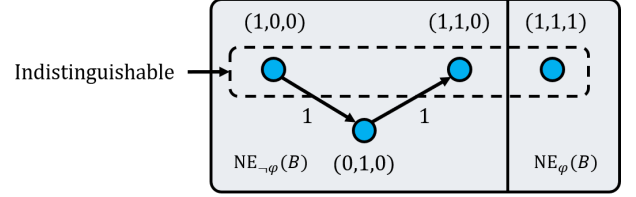| Valuation $\vec{v}$	Observation $\vec{o}$	$\vec{v} \in \text{NE}(B)$?	$\vec{v} \models \varphi$?
(0,0,0)	0	✗	✗
(0,0,1)	0	✗	✗
(0,1,0)	0	✓	✗
(0,1,1)	0	✗	✗
(1,0,0)	1	✓	✗
(1,0,1)	1	✗	✗
(1,1,0)	1	✓	✗
(1,1,1)	1	✓	✓

Figure 2: Example of a two-player costless game in which $\Phi = \{p_1, p_2, p_3\}$, $\Phi_1 = \{p_1, p_2\}$, $\Phi_2 = \{p_3\}$, $\gamma_1 = p_1 \vee p_2$, $\gamma_2 = p_3 \Rightarrow (p_1 \wedge p_2)$, $\mathcal{O} = \{p_1\}$, $\vec{o}_0 = 0$, $\vec{o}_1 = (1)$ and $\varphi = p_1 \wedge p_2 \wedge p_3$. The table on the top illustrates different properties of each strategy profile, and the diagram on the bottom depicts a deviation path which contains an observed deviation cycle.

Theorem 1. *Let B be a Boolean game, $\mathcal{O}$ an observable set, and $X \subseteq \text{INIT}(B)$ a non-empty set of initial equilibria. Then, X is eliminable if and only if there exists a deviation graph $D = (\mathcal{V}_D, E_D)$ of B with respect to $\mathcal{O}$ that satisfies the following properties: 1) For every $\vec{v} \in X$, there is some $\vec{v}' \in \mathcal{V}_D$ such that $(\vec{v}, \vec{v}') \in E_D$; and 2) Every observed deviation cycle in D involves at least two agents;*

Proof Sketch. The forward direction can be shown via proof by contrapositive, which proceeds under a case analysis. If property 1) does not hold for a deviation graph D, then there is some strategy profile in X which will not be eliminated by inducing D. If property 2) does not hold, then it is impossible to induce D, because doing so would lead to a contradiction by transitivity of the preference relation. The backward direction can be shown by constructing a contract κ that assigns rewards using a 'backward induction' procedure, which induces the deviation graph D. It can then be easily checked using the properties of D that κ eliminates X. $\square$

5.2 Contract Design for Logical Objectives

Returning to the problem of designing contracts for the guaranteed satisfaction of logical objectives, the natural question to ask is, given a Boolean game B, does there exist a contract κ such that the principal can ensure that their goal φ has been satisfied on some or every Nash equilibrium of the game B^κ? The E-Nash version of this problem is defined as follows:

E-NASH CONTRACTIBILITY:
Given: Game B, observable set $\mathcal{O}$, formula φ.
Question: Does there exist a contract κ such that for some $\vec{o} \in \mathbb{B}^\mathcal{O}$ we have $\text{ENV}(B^\kappa, \vec{o}, \varphi) \neq \emptyset$?

Note that in the problem of contract design, the principal is not given an observation to begin with, but must consider *all possible observations* that are consistent with at least one

Nash equilibrium of a given Boolean game – since the agents' utilities are affected by the imposed contract, their strategies will likewise be influenced by the incentives introduced by the contract and hence, this component must be specified *before* agents in the game select their strategies. We have found that this problem is no harder than the special case where the principal has full observability within the game [Wooldridge *et al.*, 2013]:

Theorem 2. E-NASH CONTRACTIBILITY *is* Σ_2^p-*complete.*

Proof. For membership in Σ_2^p, first observe that the answer to E-Nash Contractibility is "yes" if and only if it is the case that $\vec{v} \in \text{IND}(B, \mathcal{O}) \cap \text{NE}_\varphi(B) \neq \emptyset$, that is, if there exists an inducible equilibrium that satisfies φ. Making use of the characterisation of inducible equilibria in Proposition 4, we can guess a strategy profile $\vec{v}$ and then check whether 1) $\vec{v} \in \text{NE}_\varphi(B)$; and 2) For all agents $i \in N$, and all choices $v_i' \in V_i$ such that $(\vec{v}_{-i}, v_i') =_\mathcal{O} \vec{v}$ and $i \in W(\vec{v}) \Leftrightarrow i \in W(\vec{v}_{-i}, v_i')$, we have that $c_i(\vec{v}_{-i}, v_i') \geq c_i(\vec{v})$. Checking the first condition is a coNP problem [Wooldridge *et al.*, 2013]. Moreover, checking the second condition is also in coNP – simply guess an alternative choice $v_i' \in V_i$ for each agent $i \in N$ and then checking whether $(\vec{v}_{-i}, v_i') =_\mathcal{O} \vec{v}$, $\vec{v} \leftrightarrows (\vec{v}_{-i}, v_i')$, and $c_i(\vec{v}_{-i}, v_i') < c_i(\vec{v})$ can be done in polynomial time. If the answer is "yes", then condition 2 is not satisfied. Thus, the overall procedure is in the Σ_2^p complexity class.

Hardness follows again from the fact that the problem of checking whether or not a Nash equilibrium exists in a cost-free Boolean game is Σ_2^p-complete [Wooldridge *et al.*, 2013]. Given an instance of such a problem, we can check for E-Nash contractibility of the formula $\varphi = \top$ in the same game, where the principal is able to observe all of the variables (i.e., $\mathcal{O} = \Phi$). The answer to the E-Nash contractibility problem will be "yes" if and only if a Nash equilibrium exists in the corresponding cost-free Boolean game, by Proposition 2. $\square$

Finally, we introduce and settle the complexity of the universal counterpart to the E-Nash Contractibility problem:

A-NASH CONTRACTIBILITY:
Given: Game B, observable set $\mathcal{O}$, formula φ.
Question: Is there a contract κ s.t. $\text{NE}(B^\kappa) \neq \emptyset$ and for all observations $\vec{o} \in \mathbb{B}^\mathcal{O}$, if $\text{CONS}(B^\kappa, \vec{o}) \neq \emptyset$, then $\text{CONS}(B^\kappa, \vec{o}) \subseteq \text{NE}_\varphi(B^\kappa)$?

With this definition in place, we can then show the following result, which sharply contrasts with results appearing in [Wooldridge *et al.*, 2013]:[4]

Theorem 3. A-NASH CONTRACTIBILITY *is* Σ_3^p-*complete.*

Proof Sketch. For membership, note that the A-Nash Contractibility problem is equivalent to deciding the following statement: there are $\kappa \in \mathcal{K}$ and $\vec{v} \in V$ such that for all $\vec{v}' \in V$ we have $(\vec{v} \in \text{NE}(B^\kappa)) \wedge (\vec{v}' \in \text{NE}(B^\kappa) \Rightarrow \vec{v}' \models \varphi)$, which is a Σ_3^p predicate, since there are at most an exponential number of polynomial-sized contracts to be checked[5], and checking

---

[4]This contradicts Proposition 14 in [Wooldridge *et al.*, 2013]. Private communication with the authors has confirmed that there was an error in the original publication [Wooldridge *et al.*, 2013].

[5]More details in the arXiv version [Hyland *et al.*, 2023].

whether $\vec{v} \in \text{NE}(B^\kappa)$ for some $\vec{v} \in V$ and $\kappa \in \mathcal{K}$ is coNP-complete.

For hardness, we reduce from QSAT$_3$, which is known to be Σ_3^p-complete [Papadimitriou, 1994]. Suppose that we have an instance of QSAT$_3$, which is given by a Boolean formula φ over a set X of Boolean variables and a partition of X into three sets X_1, X_2, X_3. The question is to decide whether $\exists X_1 \forall X_2 \exists X_3 \; \varphi$. We transform this into an instance of A-Nash Contractibility by defining a three-agent game $B = (\{1, 2, 3\}, \Phi, (\Phi_i)_{i \in N}, (\gamma_i)_{i \in N}, (c_i)_{i \in N})$, where: $\Phi = X \cup \{p, q\}$, $\Phi_1 = X_1$, $\Phi_2 = X_2 \cup \{p\}$, and $\Phi_3 = X_3 \cup \{q\}$; $\gamma_1 = \top$; $\gamma_2 = \neg\varphi \vee (p \leftrightarrow q)$; $\gamma_3 = \varphi \vee \neg(p \leftrightarrow q)$; For all $\vec{v} \in V$ and $i \in \{1, 2\}$, we have $c_i(\vec{v}) = 0$; For all $\vec{v} \in V$, we have $c_3(\vec{v}) = 1$ if $\vec{v} \nvDash \varphi$ and is 0 otherwise. The objective that the principal wishes to implement is simply φ, and the observable set is given by $\mathcal{O} = X_1$. The proof is completed by verifying that the answer to an instance of QSAT$_3$ is "yes" if and only if the answer to A-Nash Contractibility in the above construction is also "yes." $\square$

This result formally establishes the intuitive notion that the task of designing incentives to eliminate *all* undesirable behaviours in a multi-agent system is, in general, significantly more challenging than creating space for a desirable outcome to be chosen among many possible outcomes in the game.

6 Concluding Remarks

Boolean games provide a very natural model in which to draw a connection between two previously separate yet closely related areas of study: moral hazard problems (from Economics) and rational verification (from AI verification). By extending the moral hazard problem to a qualitative setting through the use of Boolean variables and propositional logic goals, this framework provides a method for expressing relationships between discrete tasks which may require a threshold of resources (costs) to complete, rather than being tied to a continuous level of effort. Our work develops this connection with a number of contributions: the formulation of a model of multi-agent moral hazard problems with combined qualitative and quantitative preferences, a characterisation of when equilibria can be induced or eliminated, and results settling the computational complexity of the verifiability and contractibility problems associated with these model of games.

As this article and previous work on moral hazard problems demonstrate, the presence of hidden actions limits a principal's ability to design contracts that successfully align agents' local incentives with the principal's higher-level objectives. This study presents a model which opens up many possible extensions and questions. Further work may refine the model in several ways by introducing an explicit *quantitative utility function for the principal* which is decreasing in contract payments, adding *randomness* to observations, allowing agents to *report* some of their actions, requiring that contracts be *individually rational*, and letting the principal *decide which observations* to make under the assumption that observations are costly. Such extensions would provide further insights into the computational aspects of incentive design in moral hazard settings, which are important concerns in AI research.

Acknowledgements

References

[Ågotnes *et al.*, 2013] Thomas Ågotnes, Paul Harrenstein, Wiebe Van Der Hoek, and Michael Wooldridge. Verifiable equilibria in boolean games. In *IJCAI*, 2013.

[Alon *et al.*, 2021] Tal Alon, Paul Dütting, and Inbal Talgam-Cohen. Contracts with private cost per unit-of-effort. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 52–69, 2021.

[Arrow, 1971] Kenneth J Arrow. Insurance, risk and resource allocation. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*, 1971.

[Babaioff *et al.*, 2006] Moshe Babaioff, Michal Feldman, and Noam Nisan. Combinatorial agency. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 18–28, 2006.

[Bloem *et al.*, 2009] Roderick Bloem, Krishnendu Chatterjee, Thomas A Henzinger, and Barbara Jobstmann. Better quality in synthesis through quantitative objectives. In *Computer Aided Verification: 21st International Conference, CAV 2009, Grenoble, France, June 26-July 2, 2009. Proceedings 21*, pages 140–156. Springer, 2009.

[Bulling and Goranko, 2022] Nils Bulling and Valentin Goranko. Combining quantitative and qualitative reasoning in concurrent multi-player games. *Autonomous Agents and Multi-Agent Systems*, 36:1–33, 2022.

[Castiglioni *et al.*, 2023] Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Multi-agent contract design: How to commission multiple agents with individual outcome. *arXiv preprint arXiv:2301.13654*, 2023.

[Chatterjee *et al.*, 2005] Krishnendu Chatterjee, Thomas A Henzinger, and Marcin Jurdzinski. Mean-payoff parity games. In *20th Annual IEEE Symposium on Logic in Computer Science (LICS'05)*, pages 178–187. IEEE, 2005.

[Che and Yoo, 2001] Yeon-Koo Che and Seung-Weon Yoo. Optimal incentives for teams. *American Economic Review*, 91(3):525–541, 2001.

[Clercq *et al.*, 2014] Sofie De Clercq, Steven Schockaert, Martine De Cock, and Ann Nowé. Possibilistic boolean games: Strategic reasoning under incomplete information. In *European Workshop on Logics in Artificial Intelligence*, pages 196–209. Springer, 2014.

[Dütting *et al.*, 2019] Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 369–387, 2019.

[Dütting *et al.*, 2021] Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. The complexity of contracts. *SIAM Journal on Computing*, 50(1):211–254, 2021.

[Dütting *et al.*, 2022] Paul Dütting, Tomer Ezra, Michal Feldman, and Thomas Kesselheim. Multi-agent contracts. *arXiv preprint arXiv:2211.05434*, 2022.

[Dütting *et al.*, 2023] Paul Dütting, Michal Feldman, and Daniel Peretz. Ambiguous contracts. *arXiv preprint arXiv:2302.07621*, 2023.

[Emek and Feldman, 2012] Yuval Emek and Michal Feldman. Computing optimal contracts in combinatorial agencies. *Theoretical Computer Science*, 452:56–74, 2012.

[Gan *et al.*, 2022] Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. Optimal coordination in generalized principal-agent problems: A revisit and extensions. *arXiv preprint arXiv:2209.01146*, 2022.

[Georgiadis, 2023] George Georgiadis. Contracting with moral hazard: A review of theory & empirics. *The Elgar Encyclopedia on the Economics of Competition and Regulation*, 2023.

[Grossman and Hart, 1992] Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. In *Foundations of insurance economics*, pages 302–340. Springer, 1992.

[Guruganesh *et al.*, 2021] Guru Guruganesh, Jon Schneider, and Joshua R Wang. Contracts under moral hazard and adverse selection. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 563–582, 2021.

[Gutierrez *et al.*, 2017] Julian Gutierrez, Aniello Murano, Giuseppe Perelli, Sasha Rubin, and Michael Wooldridge. Nash equilibria in concurrent games with lexicographic preferences. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, page 1067–1073, 2017.

[Gutierrez *et al.*, 2019] Julian Gutierrez, Muhammad Najib, Giuseppe Perelli, and Michael Wooldridge. Equilibrium Design for Concurrent Games. In Wan Fokkink and Rob van Glabbeek, editors, *30th International Conference on Concurrency Theory (CONCUR 2019)*, volume 140 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 22:1–22:16, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Gutierrez *et al.*, 2021] Julian Gutierrez, Aniello Murano, Giuseppe Perelli, Sasha Rubin, Thomas Steeples, and Michael Wooldridge. Equilibria for games with combined qualitative and quantitative objectives. *Acta Informatica*, 58(6):585–610, 2021.

[Harrenstein *et al.*, 2001] Paul B. Harrenstein, Wiebe van der Hoek, J.-J.Ch. Meyer, and C. Witteveen. Boolean games. In *TARK*, pages 287–298, 2001.

[Harrenstein *et al.*, 2014] Paul B. Harrenstein, Paolo Turrini, and Michael J. Wooldridge. Hard and soft equilibria in boolean games. In *AAMAS*, volume 14, pages 5–9, 2014.

[Harrenstein *et al.*, 2017] Paul B. Harrenstein, Paolo Turrini, and Michael Wooldridge. Characterising the manipulability of boolean games. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.

[Holmström, 1979] Bengt Holmström. Moral hazard and observability. *The Bell journal of economics*, pages 74–91, 1979.

[Holmstrom, 1982] Bengt Holmstrom. Moral hazard in teams. *The Bell journal of economics*, pages 324–340, 1982.

[Hyland *et al.*, 2023] David Hyland, Julian Gutierrez, and Michael Wooldridge. Principal-agent boolean games. *arXiv preprint arXiv:2305.10334*, 2023.

[Itoh, 1991] Hideshi Itoh. Incentives to help in multi-agent situations. *Econometrica: Journal of the Econometric Society*, pages 611–636, 1991.

[Levit *et al.*, 2019] Vadim Levit, Zohar Komarovsky, Tal Grinshpoun, Ana LC Bazzan, and Amnon Meisels. Incentive-based search for equilibria in boolean games. *Constraints*, 24(3):288–319, 2019.

[Maubert *et al.*, 2021] Bastien Maubert, Munyque Mittelmann, Aniello Murano, and Laurent Perrussel. Strategic reasoning in automated mechanism design. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 18, pages 487–496, 2021.

[Michael and William, 1976] C Jensen Michael and H Meckling William. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3(4):305–360, 1976.

[Mittelmann *et al.*, 2022] Munyque Mittelmann, Bastien Maubert, Aniello Murano, and Laurent Perrussel. Automated synthesis of mechanisms. In *31st International Joint Conference on Artificial Intelligence (IJCAI-22)*, pages 426–432. International Joint Conferences on Artificial Intelligence Organization, 2022.

[Narayanaswami *et al.*, 2022] Sai Kiran Narayanaswami, Swarat Chaudhuri, Moshe Vardi, and Peter Stone. Automating mechanism design with program synthesis. *Proc. of the Adaptive and Learning Agents Workshop (ALA 2022)*, 2022.

[Nisan and Ronen, 1999] Noam Nisan and Amir Ronen. Algorithmic mechanism design. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 129–140, 1999.

[Papadimitriou, 1994] Christos. H. Papadimitriou. *Computational Complexity*. AW, 1994.

[Pauly, 1968] Mark V Pauly. The economics of moral hazard: comment. *The american economic review*, 58(3):531–537, 1968.

[Shavell, 1979] Steven Shavell. Risk sharing and incentives in the principal and agent relationship. *The Bell Journal of Economics*, pages 55–73, 1979.

[Smith, 1980] Reid. G. Smith. *A Framework for Distributed Problem Solving*. UMI R. Press, 1980.

[Wooldridge *et al.*, 2013] Michael Wooldridge, Ulle Endriss, Sarit Kraus, and Jérôme Lang. Incentive engineering for boolean games. *Artificial Intelligence*, 195:418–439, 2013.