Incentive Engineering for Boolean Games*

Ulle Endriss[1] **Sarit Kraus**[2] **Jérôme Lang**[3] **Michael Wooldridge**[4]

[1]University of Amsterdam, The Netherlands (ulle.endriss@uva.nl)
[2]Bar Ilan University, Israel (sarit@cs.biu.ac.il)
[3]Université Paris-Dauphine, France (lang@lamsade.dauphine.fr)
[4]University of Liverpool, United Kingdom (mjw@liv.ac.uk)

Abstract

We investigate the problem of influencing the preferences of players within a *Boolean game* so that, if all players act rationally, certain desirable outcomes will result. The way in which we influence preferences is by overlaying games with *taxation schemes*.

In a Boolean game, each player has unique control of a set of Boolean variables, and the choices available to the player correspond to the possible assignments that may be made to these variables. Each player also has a goal, represented by a Boolean formula, that they desire to see satisfied. Whether or not a player's goal is satisfied will depend both on their own choices and on the choices of others, which gives Boolean games their strategic character. We extend this basic framework by introducing an external principal who is able to levy a taxation scheme on the game, which imposes a cost on every possible action that a player can choose. By designing a taxation scheme appropriately, it is possible to perturb the preferences of the players, so that they are incentivised to choose some equilibrium that would not otherwise be chosen. After motivating and formally presenting our model, we explore some issues surrounding it, including the complexity of finding a taxation scheme that implements some socially desirable outcome, and then discuss desirable properties of taxation schemes.

1 Introduction

Our goal is to investigate the possibility of influencing the behaviour of rational players in a game towards certain outcomes by *providing incentives* for them to act in certain ways. If we look to the real world, we see two forms of incentives that are typically used by governments and other organisations in order to influence human behaviour: we can call them "carrots" and "sticks". "Carrots" provide positive incentives, by rewarding players who act in the desired way, while "sticks" penalise undesirable behaviour. One of the most common incentive mechanisms found in human societies is taxation. Taxation is frequently used to incentivise behaviours. For example, a government might tax car driving in order to encourage the use of environmentally friendly public transport; or it might tax cigarettes in order to discourage smoking. Of course, as well as incentivising behaviour, taxation is also used by governments to raise revenue, typically with the intention that this revenue is then used to fund socially desirable projects (education, healthcare, etc).

In the present paper, we study the design of taxation schemes for incentivising behaviours in multi-agent systems. The setting for our study is the domain of *Boolean games* [5; 1; 3]. Boolean games are a natural, expressive, and compact class of games, based on propositional logic. Boolean games were introduced in [5], and their computational and logical properties have subsequently been studied by several researchers [1; 3]. In such a game, each agent i is assumed to have a goal, represented as a propositional formula γ_i over some set of variables Φ. In addition, each agent i is allocated some subset Φ_i of the variables Φ, with the idea being that the variables Φ_i are under the unique control of agent i. The choices, or strategies, available to i correspond to all the possible allocations of truth or falsity to the variables Φ_i. An agent will try to choose an allocation so as to satisfy its goal γ_i. Strategic concerns arise because whether i's goal is in fact satisfied will depend on the choices made by others.

We introduce the idea of imposing taxation schemes on Boolean games, so that a player's possible choices are taxed in different ways. Taxation schemes are designed by an agent external to the game known as the *principal*. The ability to impose taxation schemes enables the principal to *perturb the preferences of the players in certain ways*: all other things being equal, an agent will prefer to make a choice that minimises taxes. As discussed above, the principal is assumed to be introducing a taxation scheme so as to incentivise agents to achieve a certain desirable outcome; or to incentivise agents to rule out certain undesirable outcomes. We represent the outcome that the principal desires to achieve via a propositional formula Υ: thus, the idea is that the principal will impose a taxation scheme so that agents are rationally incentivised to make individual choices so as to collectively

*This paper is an invited contribution for the IJCAI-2011 "Best Papers" track. It is an adapted and somewhat simplified version of the paper *Designing Incentives for Boolean Games*, which was accepted for the AAMAS-2011 conference and shortlisted for the best paper prize at this conference. We refer the reader to this parent paper for further technical details, proofs, and discussion.

satisfy Υ. However, a fundamentally important assumption in what follows is that taxes do not give us absolute control over an agent's preferences. To assume that we were able to completely control an agent's preferences by imposing taxes would be unrealistic: to pick a perhaps rather morbid and slightly tongue in cheek example, no matter how much you propose to tax me, I would still choose to achieve my goal of being alive rather than otherwise. If we *did* have complete control over agents' preferences through taxation, then the problems we consider in this paper would indeed be rather trivial. In our setting specifically, it is assumed that no matter what the level of taxes, *an agent would still prefer to have its goal achieved than not*. This imposes a fundamental limit on the extent to which an agent's preferences can be perturbed by taxation.

We begin in the following section by introducing the model of Boolean games that we use throughout the remainder of the paper. We then introduce taxation schemes and the *incentive design problem* – the problem of designing taxation schemes so that a certain objective Υ is satisfied in equilibrium. After investigating some issues around the incentive design problem, we go on to consider possible desirable properties of taxation schemes (such as minimising the total tax burden). We conclude with a discussion and future work.

2 Boolean Games

Propositional Logic: Throughout the paper, we make use of classical propositional logic, and for completeness, we thus begin by recalling the technical framework of this logic. Let $\mathbb{B} = \{\top, \bot\}$ be the set of Boolean truth values, with "$\top$" being truth and "$\bot$" being falsity. We will abuse notation a little by using $\top$ and $\bot$ to denote both the syntactic constants for truth and falsity respectively, as well as their semantic counterparts (i.e., the respective truth values). Let $\Phi = \{p, q, \ldots\}$ be a (finite, fixed, non-empty) vocabulary of Boolean variables, and let $\mathcal{L}$ denote the set of (well-formed) formulae of propositional logic over Φ, constructed using the conventional Boolean operators ("$\wedge$", "$\vee$", "$\rightarrow$", "$\leftrightarrow$", and "$\neg$"), as well as the truth constants "$\top$" and "$\bot$". We assume a conventional semantic consequence relation "$\models$" for propositional logic. A *valuation* is a total function $v : \Phi \rightarrow \mathbb{B}$, assigning truth or falsity to every Boolean variable. We write $v \models \varphi$ to mean that φ is true under, or satisfied by, valuation v, where the satisfaction relation "$\models$" is defined in the standard way. Let $\mathcal{V}$ denote the set of all valuations over Φ.

We write $\models \varphi$ to mean that φ is a tautology, i.e., is satisfied by every valuation. We denote the fact that formulae $\varphi, \psi \in \mathcal{L}$ are logically equivalent by $\varphi \Leftrightarrow \psi$; thus $\varphi \Leftrightarrow \psi$ means that $\models \varphi \leftrightarrow \psi$. Note that "$\Leftrightarrow$" is a meta-language relation symbol, which should not be confused with the object-language bi-conditional operator "$\leftrightarrow$".

Agents, Goals, and Controlled Variables: The games we consider are populated by a set $Ag = \{1, \ldots, n\}$ of *agents* – the players of the game. Each agent is assumed to have a *goal*, characterised by an $\mathcal{L}$-formula: we write γ_i to denote the goal of agent $i \in Ag$. Each agent $i \in Ag$ *controls* a (possibly empty) subset Φ_i of the overall set of Boolean variables (cf. [10]). By "control", we mean that i has the unique ability within the game to set the value (either $\top$ or $\bot$) of each variable $p \in \Phi_i$. We will require that $\Phi_1, \ldots, \Phi_n$ forms a partition of Φ, i.e., every variable is controlled by some agent and no variable is controlled by more than one agent ($\Phi_i \cap \Phi_j = \emptyset$ for $i \neq j$). Where $i \in Ag$, a *choice* for agent i is defined by a function $v_i : \Phi_i \rightarrow \mathbb{B}$, i.e., an allocation of truth or falsity to all the variables under i's control. Let $\mathcal{V}_i$ denote the set of choices for agent i. The intuitive interpretation we give to $\mathcal{V}_i$ is that it defines the *actions* or *strategies* available to agent i; the *choices* available to the agent.

An *outcome*, $(v_1, \ldots, v_n) \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_n$, is a collection of choices, one for each agent. Clearly, every outcome uniquely defines a valuation, and we will often think of outcomes as valuations, for example writing $(v_1, \ldots, v_n) \models \varphi$ to mean that the valuation defined by the outcome $(v_1, \ldots, v_n)$ satisfies formula $\varphi \in \mathcal{L}$. Let $\varphi_{(v_1,\ldots,v_n)}$ denote the formula that uniquely characterises the outcome $(v_1, \ldots, v_n)$:

$$\varphi_{(v_1,\ldots,v_n)} = \bigwedge_{\substack{p \in \Phi: \\ (v_1,\ldots,v_n) \models p}} p \quad \wedge \quad \bigwedge_{\substack{q \in \Phi: \\ (v_1,\ldots,v_n) \not\models q}} \neg q$$

Let $succ(v_1, \ldots, v_n)$ denote the set of agents who have their goal achieved by outcome $(v_1, \ldots, v_n)$, i.e.,:

$$succ(v_1, \ldots, v_n) = \{i \in Ag \mid (v_1, \ldots, v_n) \models \gamma_i\}.$$

Costs: Intuitively, the actions available to agents correspond to setting variables true or false. We assume that these actions have *costs*, defined by a *cost function* $c : \Phi \times \mathbb{B} \rightarrow \mathbb{R}_{\geq}$, so that $c(p, b)$ is the marginal cost of assigning the value $b \in \mathbb{B}$ to variable $p \in \Phi$.

This notion of a cost function represents an obvious generalisation of previous presentations of Boolean games: costs were not considered in the original presentation of Boolean games [5; 1], and while costs were introduced in [3], it was assumed that only the action of setting a variable to $\top$ would incur a cost. In fact, as we discuss in the parent paper, costs are, in a technical sense, not required in our framework; we can capture the key strategic issues at stake without them. This is because we can "simulate" marginal costs with taxes. However, it is natural from the point of view of modelling to have costs for actions, and to think about costs as being imposed from within the game, and taxes, (defined below), as being imposed from without.

Boolean Games: Collecting these components together, a *Boolean game*, G, is a $(2n + 3)$-tuple:

$$G = \langle Ag, \Phi, c, \gamma_1, \ldots, \gamma_n, \Phi_1, \ldots, \Phi_n \rangle,$$

where $Ag = \{1, \ldots, n\}$ is a set of agents, $\Phi = \{p, q, \ldots\}$ is a finite set of Boolean variables, $c : \Phi \times \mathbb{B} \rightarrow \mathbb{R}_{\geq}$ is a cost function, $\gamma_i \in \mathcal{L}$ is the goal of agent $i \in Ag$, and $\Phi_1, \ldots, \Phi_n$ is a partition of Φ over Ag, with the intended interpretation that Φ_i is the set of Boolean variables under the unique control of $i \in Ag$.

When playing a Boolean game, the primary aim of an agent i will be to choose an assignment of values for the variables Φ_i under its control so as to satisfy its goal γ_i. The difficulty is that γ_i may contain variables controlled by other agents

$j \neq i$, who will also be trying to choose values for their variables Φ_j so as to get their goals satisfied; and their goals in turn may be dependent on the variables Φ_i. Note that if an agent has multiple ways of gettings its goal achieved, then it will prefer to choose one that minimises costs; and if an agent cannot get its goal achieved, then it simply chooses to minimise costs. These considerations are what give Boolean games their strategic character. For the moment, we will postpone the formal definition of the utility functions and preferences associated with our games.

Example 1 *Consider a simple example, to illustrate the general setup of Boolean games and the problem we consider in this paper. Suppose we have a game with two players,* $Ag = \{1, 2\}$. *There are just three variables in the game:* p, q *and* r, *i.e.,* $\Phi = \{p, q, r\}$. *Player* 1 *controls* p *(so* $\Phi_1 = \{p\}$*), while player* 2 *controls* q *and* r *(i.e.,* $\Phi_2 = \{q, r\}$*). All costs are* 0*. Now, suppose the goal formulae* γ_i *for our players are defined as follows:*

$$\begin{array}{lcl} \gamma_1 & = & q \\ \gamma_2 & = & q \vee r \end{array}$$

Notice that player 1 *is completely dependent on player* 2 *for the achievement of his goal, in the sense that, for player* 1 *to have his goal achieved, player* 2 *must set* $q = \top$. *However, player* 2 *is* not *dependent on player* 1*: he is in the fortunate position of being able to achieve his goal entirely through his own actions, irrespective of what others do. He can either set* $q = \top$ *or* $r = \top$, *and his goal will be achieved. What will the players do? Well, in this case, the game can be seen as having a happy outcome: player* 2 *can set* $q = \top$, *and both agents will get their goal satisfied at no cost. Although we have not yet formally defined the notion, we can informally see that this outcome forms an equilibrium, in the sense that neither player has any incentive to do anything else.*

Now let us change the game a little. Suppose the cost for player 2 *of setting* $q = \top$ *is* 10*, while the cost of setting* $q = \bot$ *is* 0*, and that all other costs in the game are* 0*. Here, although player* 2 can *choose an action that satisfies the goal of player* 1*, he will not rationally choose it, because it is more expensive. Player* 2 *would prefer to set* $r = \top$ *than to set* $q = \top$, *because this way he would get his goal achieved at no cost. However, by doing so, player* 1 *is left without his goal being satisfied, and with no way to satisfy his goal. Now, it could be argued that the outcome here is* socially undesirable*, because it would be possible for* both *players to get their goal achieved. Our idea in the present paper is to provide incentives for player* 2 *so that he* will *choose the more socially desirable outcome in which both players get their goal satisfied. The incentives we study are in the form of taxes: we tax player* 2*'s actions so that setting* $q = \top$ *is cheaper than setting* $r = \top$, *and so the socially desirable outcome results. This might seem tough on player* 2*, but notice that he still gets his goal achieved. And in fact, as we will see below, there are limits to the kind of behaviour we can incentivise by taxes. In a formal sense, to be defined below, there is nothing we can do that would induce player* 2 *to set* both q *and* r *to* $\bot$, *since this would result in his goal being unsatisfied.*

3 Designing Incentives

We can now describe in more detail the overall problem that we consider in the remainder of the paper. Imagine a society populated by agents Ag, with each agent $i \in Ag$ having a goal $\gamma_i \in \mathcal{L}$ and actions corresponding to valuations to Φ_i. We assume an external *principal* has some goal $\Upsilon \in \mathcal{L}$ that it wants the society to achieve, and to this end, wants to incentivise the agents Ag to act collectively so as to bring about Υ. Incentives in our model are provided by *taxation schemes*.

Taxation Schemes: A taxation scheme defines additional (imposed) costs on actions, over and above those given by the marginal cost function c. While the cost function c is fixed and immutable for any given Boolean game, the principal is assumed to be at liberty to levy taxes as they see fit. Agents will seek to minimise their overall costs, and so by assigning different levels of taxation to different actions, the principal can incentivise agents away from performing some actions and towards performing others; if the principal designs the taxation scheme correctly, then agents are incentivised to choose valuations $(v_1, \ldots, v_n)$ so as to satisfy Υ (i.e., so that $(v_1, \ldots, v_n) \models \Upsilon$).

We model a taxation scheme as a function $\tau : \Phi \times \mathbb{B} \rightarrow \mathbb{R}_{\geq}$, where the intended interpretation is that $\tau(p, b)$ is the tax that would be levied on the agent controlling p if the value b was assigned to the Boolean variable p. The *total* tax paid by an agent i in choosing a valuation $v_i \in \mathcal{V}_i$ will be $\sum_{p \in \Phi_i} \tau(p, v_i(p))$.

We let τ_0 denote the taxation scheme that applies no taxes to any choice, i.e., $\forall x \in \Phi$ and $b \in \mathbb{B}$, $\tau_0(x, b) = 0$. Let $\mathcal{T}(G)$ denote the set of taxation schemes over G. We make one technical assumption in what follows, relating to the space requirements for taxation schemes in $\mathcal{T}(G)$. Unless otherwise stated explicitly, we will assume that we are restricting our attention to taxation schemes whose values can be represented with a space requirement that is bounded by a polynomial in the size of the game. This seems a reasonable requirement: realistically, taxation schemes requiring space exponential in the size of the game at hand could not be manipulated. It is important to note that this requirement relates to the *space requirements for taxes*, and not to the *size of taxes themselves*: for a polynomial function $f : \mathbb{N} \rightarrow \mathbb{N}$, the value $2^{f(n)}$ can be represented using only a polynomial number of bits (i.e., $f(n)$ bits).

Utilities and Preferences: One important assumption we make is that while taxation schemes can influence the decision making of rational agents, they cannot, ultimately, change the goals of an agent. That is, if an agent has a chance to achieve its goal, it will take it, no matter what the taxation incentives are to do otherwise. To understand this point, and to see formally how incentives work, we need to formally define the utility functions for agents, and for this we require some further auxiliary definitions. First, with a slight abuse of notation, we extend cost and taxation functions to partial valuations as follows:

$$c_i(v_i) = \sum_{p \in \Phi_i} c(p, v_i(p))$$

$$\tau_i(v_i) = \sum_{p \in \Phi_i} \tau(p, v_i(p))$$

Next, let v_i^e denote the most expensive possible course of action for agent i:

$$v_i^e \in \arg\max_{v_i \in \mathcal{V}_i} (c_i(v_i) + \tau_i(v_i)).$$

Let μ_i denote the cost to i of its most expensive course of action:

$$\mu_i = c_i(v_i^e) + \tau_i(v_i^e).$$

Given these definitions, we define the *utility* to agent i of an outcome $(v_1, \ldots, v_n)$, as follows:

$$u_i(v_1, \ldots, v_n) = \begin{cases} 1 + \mu_i - (c_i(v_i) + \tau_i(v_i)) & \text{if } (v_1, \ldots, v_n) \models \gamma_i \\ -(c_i(v_i) + \tau_i(v_i)) & \text{otherwise.} \end{cases}$$

Thus utility for agent i will range from $1 + \mu_i$ (the best outcome for i, where it gets its goal achieved by performing actions that have no tax or other cost) down to $-\mu_i$ (where i does not get its goal achieved but makes its most expensive choice). This definition has the following properties:

- an agent prefers all outcomes that satisfy its goal over all those that do not satisfy it;
- between two outcomes that satisfy its goal, an agent prefers the one that minimises total expense (= marginal costs + taxes); and
- between two valuations that *do not* satisfy its goal, an agent prefers to minimise total expense.

It is important to note that while utility functions provide a convenient numeric representation of preference relations, utility is *not* transferable in our settings.

Solution Concepts: Given this formal definition of utility, we can define solution concepts in the standard game-theoretic way [9]. In this paper, we focus on (pure) Nash equilibrium. (Of course, other solution concepts, such as dominant strategy equilibria, might also be considered, but for simplicity, in this paper we focus on Nash equilibria.) We say an outcome $(v_1, \ldots, v_i, \ldots, v_n)$ is a Nash equilibrium if for all agents $i \in Ag$, there is no $v_i' \in \mathcal{V}_i$ such that $u_i(v_1, \ldots, v_i', \ldots, v_n) > u_i(v_1, \ldots, v_i, \ldots, v_n)$. Let $NE(G, \tau)$ denote the set of all Nash equilibria of the game G with taxation scheme τ.

Before proceeding, let us consider some properties of Nash equilibrium outcomes. First, observe that an unsuccessful agent will choose a least cost course of action in any Nash equilibrium.

Proposition 1 *Suppose* $(v_1^*, \ldots, v_i^*, \ldots, v_n^*) \in NE(G, \tau)$ *is such that* $i \notin succ(v_1^*, \ldots, v_i^*, \ldots, v_n^*)$. *Then*

$$v_i^* \in \arg\min_{v_i \in \mathcal{V}_i} c_i(v_i) + \tau_i(v_i)$$

The following is an obvious decision problem:

NASH OUTCOME VERIFICATION:
Instance: Boolean game G, taxation scheme τ, and outcome $(v_1, \ldots, v_n)$.
Question: Is $(v_1, \ldots, v_n) \in NE(G, \tau)$?

Proposition 2 NASH OUTCOME VERIFICATION *is co-*NP*-complete, even for two player games with* $\tau = \tau_0$ *and where* c *assigns no costs.*

Incentive Design: We now come to the main problems that we consider in the remainder of the paper. Suppose we have an agent, which we will call the principal, who is external to a game G. The principal is at liberty to impose taxation schemes on the game G. It will not do this for no reason, however: it does it because it wants to provide incentives for the agents in G to choose certain collective outcomes. Specifically, the principal wants to incentivise the players in G to choose rationally a collective outcome that satisfies an *objective*, which is represented as a propositional formula Υ over the variables Φ of G. We refer to this general problem – trying to find a taxation scheme that will incentivise players to choose rationally a collective outcome that satisfies a propositional formula Υ – as the *implementation problem*. It inherits concepts from the theory of Nash implementation in mechanism design [6], although our use of Boolean games, taxation schemes, and propositional formulae to represent objectives is quite different.

3.1 Weak Implementation

Let $\mathcal{WI}(G, \Upsilon)$ denote the set of taxation schemes over G that satisfy a propositional objective Υ in at least one Nash equilibrium outcome:

$$\mathcal{WI}(G, \Upsilon) = \{\tau \in \mathcal{T}(G) \mid \exists (v_1, \ldots, v_n) \in NE(G, \tau) \text{ s.t. } (v_1, \ldots, v_n) \models \Upsilon\}.$$

Given this definition, we can state the first basic decision problem that we consider in the remainder of the paper:

WEAK IMPLEMENTATION:
Instance: Boolean game G and objective $\Upsilon \in \mathcal{L}$.
Question: Is it the case that $\mathcal{WI}(G, \Upsilon) \neq \emptyset$?

If the answer to the WEAK IMPLEMENTATION problem (G, Υ) is "yes", then we say that Υ *can be weakly implemented in Nash equilibrium* (or simply: Υ can be weakly implemented in G). Let us see an example.

Example 2 *Define a game* G *as follows:* $Ag = \{1, 2\}$, $\Phi = \{p_1, p_2\}$, $\Phi_i = \{p_i\}$, $\gamma_1 = p_1$, $\gamma_2 = \neg p_1 \wedge \neg p_2$, $c(p_1, b) = 0$ *for all* $b \in \mathbb{B}$, *while* $c(p_2, \top) = 1$ *and* $c(p_2, \bot) = 0$. *Define an objective* $\Upsilon = p_1 \wedge p_2$. *Now, without any taxes (i.e., with taxation scheme* τ_0*), there is a single Nash equilibrium,* (v_1^*, v_2^*), *which satisfies* $p_1 \wedge \neg p_2$. *Agent* 1 *gets its goal achieved, while agent* 2 *does not; and moreover* $(v_1^*, v_2^*) \not\models \Upsilon$. *However, if we adjust* τ *so that* $\tau(p_2, \bot) = 10$, *then we find a Nash equilibrium outcome* (v_1', v_2') *such that* $(v_1', v_2') \models p_1 \wedge p_2$, *i.e.,* $(v_1', v_2') \models \Upsilon$. *Here, agent* 2 *is not able to get its goal achieved, but it can, nevertheless, be incentivised by taxation to make a choice that ensures the achievement of the objective* Υ.

So, what objectives Υ can be weakly implemented? At first sight, it might appear that the satisfiability of Υ is a sufficient condition for implementability. Consider the following naive approach for constructing taxation schemes with the aim of implementing satisfiable objectives Υ:

> Find a valuation v such that $v \models \Upsilon$ (such a valuation will exist since Υ is satisfiable). Then define a taxation scheme τ such that $\tau(p, b) = 0$ if $b = v(p)$ and $\tau(p, b) = k$ otherwise, where k is an astronomically large number.

Thus, the idea is simply to make all choices other than selecting an outcome that satisfies Υ too expensive to be rational. In fact, this approach does not work, because of an important subtlety of the definition of utility. In designing a taxation scheme, the principal can perturb an agent's choices between different valuations, but it *cannot* perturb them in such a way that an agent would prefer an outcome that does not satisfy it's goal over an outcome that does. We have:

Proposition 3 *There exist instances of the* WEAK IMPLEMENTATION *problem with satisfiable objectives* Υ *that cannot be weakly implemented.*

What about tautologous objectives, i.e., objectives Υ such that $\Upsilon \Leftrightarrow \top$? Again, we might be tempted to assume that tautologies are trivially implementable. This is not in fact the case, however, as it may be that $NE(G, \tau) = \emptyset$ for all taxation schemes τ:

Proposition 4 *There exist instances of the* WEAK IMPLEMENTATION *problem with tautologous objectives* Υ *that cannot be implemented.*

Tautologous objectives might appear to be of little interest, but we argue that this is not the case. Suppose we have a game G such that $NE(G, \tau_0) = \emptyset$. Then, in its unmodified condition, this game is *unstable*: it has no equilibria. Thus, we will refer to the problem of implementing $\top$ (= checking for the existence of a taxation scheme that would ensure at least one Nash equilibrium outcome), as the STABILISATION problem. The following example illustrates STABILISATION.

Example 3 *Let* $Ag = \{1, 2, 3\}$, *with* $\varphi = \{p, q, r\}$, $\Phi_1 = \{p\}$, $\Phi_2 = \{q\}$, $\Phi_3 = \{r\}$, $\gamma_1 = \top$, $\gamma_2 = (q \wedge \neg p) \vee (q \leftrightarrow r)$, $\gamma_3 = (r \wedge \neg p) \vee \neg(q \leftrightarrow r)$, $c(p, \top) = 0$, $c(p, \bot) = 1$, *and all other costs are* 0. *For any outcome in which* $p = \bot$, *agent* 1 *would prefer to set* $p = \top$, *so no such outcome can be stable. So, consider outcomes* (v_1, v_2, v_3) *in which* $p = \top$. *Here if* $(v_1, v_2, v_3) \models q \leftrightarrow r$ *then agent* 3 *would prefer to deviate, while if* $(v_1, v_2, v_3) \not\models q \leftrightarrow r$ *then agent* 2 *would prefer to deviate. Now, consider a taxation scheme with* $\tau(p, \top) = 10$ *and* $\tau(p, \bot) = 0$ *and all other taxes are* 0. *With this scheme, the outcome in which all variables are set to* $\bot$ *is a Nash equilibrium. Hence this taxation scheme stabilises the system.*

Returning to the weak implementation problem, we can derive a *sufficient* condition for weak implementation, as follows.

Proposition 5 *For all games G and objectives* Υ, *if the formula* Υ' *is satisfiable:*

$$\Upsilon' = \Upsilon \wedge \bigwedge_{i \in Ag} \gamma_i$$

then $\mathcal{WI}(G, \Upsilon) \neq \emptyset$.

We know from [1] that the problem of checking for the existence of pure strategy Nash equilibria in cost-free Boolean games is Σ_2^p-complete. It turns out that the IMPLEMENTATION problem is no harder:

Proposition 6 *The* STABILISATION *problem is* Σ_2^p*-complete, even if taxes are* 0*-bounded. As a consequence, the* WEAK IMPLEMENTATION *problem is also* Σ_2^p*-complete.*

3.2 (Strong) Implementation

The fact that $\mathcal{WI}(G, \Upsilon) \neq \emptyset$ is good news of a kind – it tells us that we can impose a taxation scheme such that *at least one* rational (NE) outcome of the game satisfies Υ. However, it could be that there are many taxation schemes, and only one of them satisfies Υ. This motivates us to consider the *strong implementation* (or simply *implementation*) problem. Strong implementation corresponds closely to the notion of Nash implementation in the mechanism design literature [6]. Let $\mathcal{SI}(G, \Upsilon)$ denote the set of taxation schemes τ over G such that:

1. G, τ has at least one Nash equilibrium outcome;
2. all Nash equilibrium outcomes of G, τ satisfy Υ.

Formally:

$$\begin{array}{l}\mathcal{SI}(G, \Upsilon) = \\ \{\tau \in \mathcal{T}(G) \mid \\ \quad NE(G, \tau) \neq \emptyset \quad \& \\ \quad \forall (v_1, \ldots, v_n) \in NE(G, \tau) : (v_1, \ldots, v_n) \models \Upsilon\}.\end{array}$$

This gives us the following decision problem:

> IMPLEMENTATION:
> *Instance*: Boolean game G and objective $\Upsilon \in \mathcal{L}$.
> *Question*: Is it the case that $\mathcal{SI}(G, \Upsilon) \neq \emptyset$?

It turns out that strong implementation is no harder than weak implementation:

Proposition 7 IMPLEMENTATION *is* Σ_2^p*-complete.*

4 Desirable Properties of Taxation Schemes

We saw above that one simple approach to designing taxation schemes is simply to apply punatively high taxes to all undesirable actions, effectively leaving players no choice but to comply with the desires of the principal. Even allowing for the key fact that, as we noted earlier, we cannot *completely* control a player's preferences using this approach (because a player would always prefer to get their goal achieved than not, however high taxes are set), this does not seem an intuitively sensible approach in practice, because arbitrarily high taxes are *ineffecient* if a player ends up paying more than they strictly need to. So, the overall goal of the principal is to design taxation schemes so as to bring about the objective Υ, and thus the first measure of whether a taxation scheme τ succeeds will be whether it implements Υ; but we can surely think of many secondary criteria through which the desirability or otherwise of a taxation scheme to implement Υ can be evaluated. In the parent paper we investigate a number of different such criteria. Here, we will focus on just two.

The first idea we have is to design a taxation scheme that implements Υ while *imposing the lowest possible tax burden on society*. Broadly, we can think of this approach as minimising the degree of intervention of the principal in the operation of society. The function $tb(\cdots)$ gives the total tax burden of an outcome:

$$tb(v_1, \ldots, v_n) = \sum_{i \in Ag} \tau(v_i).$$

The optimal taxation scheme τ^* then satisfies:

$$\tau^* \in \arg\min_{\tau \in \mathcal{SI}(G,\Upsilon)} \max\{tb(v_1, \ldots, v_n) \mid (v_1, \ldots, v_n) \in NE(G, \tau)\}$$

It is easy to construct examples showing that minimising the total tax burden may result in socially undesirable outcomes; but such "least intervention" approaches are of course very popular in human societies.

Another desirable property of taxation schemes is that they should treat those in similar circumstances broadly the same. In the literature on taxation, this is known as *horizontal equity* [2]. One could formalise this notion in several different ways for our model, but we will focus on the following idea. In any outcome, we have two "classes" of agents: those that get their goal achieved and those that do not. Thus, when looking at the differences in taxes paid, we only compare the taxes of agents that get their goal achieved against other agents that get their goal achieved, and we only compare agents that do not get their goal achieved against other agents that do not get their goal achieved. The function $he(\cdots)$ denotes the maximum difference in tax paid between agents in the same equivalence class:

$$he(v_1, \ldots, v_n) = \max \{abs(\tau_i(v_i) - \tau_j(v_j)) \mid \{i,j\} \subseteq Ag \;\&\; (v_1, \ldots, v_n) \models \gamma_i \wedge \gamma_j\} \cup \{abs(\tau_i(v_i) - \tau_j(v_j)) \mid \{i,j\} \subseteq Ag \;\&\; (v_1, \ldots, v_n) \models \neg(\gamma_i \vee \gamma_j)\}$$

Then τ^* will denote an outcome that maximises horizontal equity (i.e., minimises the difference in taxes paid by agents in the same circumstances).

$$\tau^* \in \arg\min_{\tau \in \mathcal{SI}(G,\Upsilon)} \max\{he(v_1, \ldots, v_n) \mid (v_1, \ldots, v_n) \in NE(G, \tau)\}$$

5 Conclusions & Future Work

We have studied the use of taxation schemes to incentivise behaviours in Boolean games. We showed how a principal can perturb the preferences of agents in a Boolean game by imposing a taxation scheme, and in so doing, how it can, in certain circumstances, incentivise agents to choose outcomes to satisfy some social objective Υ, represented as a Boolean formula. However, we saw that while an agent's preferences can be perturbed, they are not completely malleable: no matter what the taxation scheme, an agent would always prefer to get its goal achieved than otherwise. This means there are limits on the extent to which preferences can be perturbed by taxation, and hence limits on what objectives Υ can be achieved. We studied a number of issues around the problem of implementing objectives Υ via taxation schemes, and also discussed the notion of equitable taxation.

Our focus in the present paper has *not* been on the design of incentive compatible mechanisms, and in this respect, our work differs from the large body of work on computational and algorithmic mechanism design [8; 4; 7]. Of course, this is not to say that incentive compatibility is not important; we are simply focussing on scenarios in which the true preferences of agents are already known and where we want to incentivise these agents to realise a range of social objectives that can be expressed in terms of a Boolean formula. We believe the results of the present paper strongly indicate that there are important and interesting theoretical and practical questions relating to non-incentive compatible taxation schemes. Future work might consider: a characterisation of the conditions under which an objective Υ can be implemented in a game G; consideration of the computation of taxation schemes τ for objectives Υ; and the use of taxation schemes to incentivise behaviour in other settings, beyond Boolean games.

Acknowledgments: This research was financially supported by the Royal Society, MOST (#3-6797), and ISF (#1357/07).

References

[1] E. Bonzon, M.-C. Lagasquie, J. Lang, and B. Zanuttini. Boolean games revisited. In *Proceedings of the Seventeenth European Conference on Artificial Intelligence (ECAI-2006)*, Riva del Garda, Italy, 2006.

[2] J. J. Cordes. Horizontal equity. In *The Encyclopedia of Taxation and Tax Policy*. Urban Institute Press, 1999.

[3] P. E. Dunne, S. Kraus, W. van der Hoek, and M. Wooldridge. Cooperative boolean games. In *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2008)*, Estoril, Portugal, 2008.

[4] E. Ephrati and J. S. Rosenschein. The Clarke tax as a consensus mechanism among automated agents. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, Anaheim, CA, 1991.

[5] P. Harrenstein, W. van der Hoek, J.-J.Ch. Meyer, and C. Witteveen. Boolean games. In J. van Benthem, editor, *Proceeding of the Eighth Conference on Theoretical Aspects of Rationality and Knowledge (TARK VIII)*, pages 287–298, Siena, Italy, 2001.

[6] E. Maskin. The theory of implementation in Nash equilibrium: A survey. MIT Department of Economics Working Paper, 1983.

[7] N. Nisan and A. Ronen. Algorithmic mechanism design. In *Proceedings of the Thirty-first Annual ACM Symposium on the Theory of Computing (STOC-99)*, pages 129–140, May 1999.

[8] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press: Cambridge, England, 2007.

[9] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press: Cambridge, MA, 1994.

[10] W. van der Hoek and M. Wooldridge. On the logic of cooperation and propositional control. *Artificial Intelligence*, 164(1-2):81–119, May 2005.