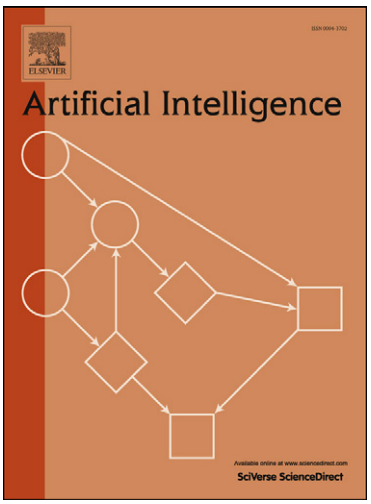
Incentive engineering for Boolean games

Michael Wooldridge, Ulle Endriss, Sarit Kraus, Jérôme Lang

Accepted Manuscript



Incentive engineering for Boolean games

Michael Wooldridge, Ulle Endriss, Sarit Kraus, Jérôme Lang

PII: S0004-3702(12)00151-8
DOI: 10.1016/j.artint.2012.11.003
Reference: ARTINT 2686

To appear in: *Artificial Intelligence*

Received date: 18 May 2012
Revised date: 13 November 2012
Accepted date: 14 November 2012

Please cite this article in press as: M. Wooldridge et al., Incentive engineering for Boolean games, *Artificial Intelligence* (2012), http://dx.doi.org/10.1016/j.artint.2012.11.003

Incentive Engineering for Boolean Games

Michael Wooldridge[+] Ulle Endriss[*] Sarit Kraus[†] Jérôme Lang[‡]

[+]Department of Computer Science
University of Oxford, United Kingdom (mjw@cs.ox.ac.uk)

[*]Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands (ulle.endriss@uva.nl)

[†]Department of Computer Science
Bar Ilan University, Israel (sarit@cs.biu.ac.il)

[‡]LAMSADE
Université Paris-Dauphine, France (lang@irit.fr)

November 19, 2012

Abstract

Boolean games are a natural, compact, and expressive class of logic-based games, in which each player exercises unique control over some set of Boolean variables, and has some logical goal formula that it desires to be achieved. A player's strategy set is the set of all possible valuations that may be made to its variables. A player's goal formula may contain variables controlled by other agents, and in this case, it must reason strategically about how best to assign values to its variables. In the present paper, we consider the possibility of overlaying Boolean games with *taxation schemes*. A taxation scheme imposes a cost on every possible assignment an agent can make. By designing a taxation scheme appropriately, it is possible to perturb the preferences of agents so that they are rationally incentivised to choose some desirable equilibrium that might not otherwise be chosen, or incentivised to rule out some undesirable equilibria. After formally presenting the model, we explore some issues surrounding it (e.g., the complexity of finding a taxation scheme that implements some desirable outcome), and then discuss possible desirable properties of taxation schemes.

1 Introduction

The computational aspects of game-theoretic mechanism design have received a great deal of attention over the past decade [23]. Particular attention has been paid to the Vickrey-Clarke-Groves (VCG) mechanism, which can be used to incentivise rational agents to truthfully report

their private preferences in settings such as combinatorial auctions [10, 22]. The key point of interest of the VCG mechanism is that, because it incentivises agents to report their preferences truthfully, it allows us to compute outcomes that maximise social welfare, which would not in general be possible if agents could benefit from misrepresenting their preferences.

Ultimately, the VCG mechanism is a *taxation scheme*. Taxation schemes are used in human societies for several purposes. First, they are used to incentivise certain desirable behaviours. For example, a government may tax private car ownership with the goal of encouraging the use of environmentally friendly public transport. Second, they are used to raise revenue, typically with the intention that this revenue is then used to fund projects such as education, healthcare, etc. And finally, of course, they may be used for a combination of these purposes. Our aim in the present paper is to study the design of taxation schemes for incentivising behaviours in multi-agent systems. It is important to note that our focus in the present paper is *not* on the design of incentive compatible (truth-telling) mechanisms, and in this key respect, our work differs from the large body of work on computational and algorithmic mechanism design [23, 10, 22]. Rather, our work considers the problem of incentivising agents whose true preferences are known towards certain courses of action. We see our work as complementing the extensive body of work on the VCG mechanism and its variants.

The setting for our study is the domain of *Boolean games*. Boolean games are a natural, expressive, and compact class of games, based on propositional logic. Boolean games were introduced by Harrenstein *et al.* [15]. Their computational and logical properties have subsequently been studied by several researchers [14, 6, 19, 8, 5, 12, 26], and [17] defines an extension of Boolean games to description logics. In a Boolean 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 associated with 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. While an agent's primary aim is to make choices so that its goal will be achieved, its secondary aim is assumed to be *minimising costs*. We assume that for every Boolean variable p and every Boolean value b, there is an associated *marginal cost*, $c(p, b)$, which would be incurred by the player controlling variable p if this player assigned to p the value b.

In the present paper, we introduce the idea of imposing *taxation schemes* on Boolean games, so that the possible choices an agent can make can be taxed in different ways. Taxation schemes are designed by an agent external to the system 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 his total expense (= marginal costs + taxes). As discussed above, the principal is assumed to be introducing a taxation scheme so as to incentivise agents to achieve a certain outcome; or to incentivise agents to rule out certain outcomes. We represent the outcome that the principal desires to achieve via a propositional logic 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 satisfy the objective 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. For example, suppose your goal is to stay alive: no matter what level of taxes are being proposed, you would surely choose to satisfy this goal 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. Such preferences can be qualified as *quasi-dichotomous*: the agent has a dichotomous utility function induced by her goal, plus a cost associated to her possible variable assignments.

Consider the following simple example, intended to illustrate the general setup of Boolean games and the problem we consider in this paper.

Example 1 *Suppose we have a game with two players,* 1 *and* 2*. There are just three variables in the game:* p, q *and* r*. Player* 1 *controls* p*, while player* 2 *controls* q *and* r*. For now, we assume all marginal costs are* 0*. Thus, a player can assign any value it desires to its variables, at no cost. This implies that a player will be indifferent between its choices on the basis of cost: all that a player is concerned about is the satisfaction of its goal.*

Now, suppose the goal formulae for our players are defined as follows:

$$\begin{array}{ll} \textit{player 1:} & q \\ \textit{player 2:} & q \vee r \end{array}$$

Thus player 1 *will be satisfied if the game results in an outcome making* q *true, while player* 2 *has a slightly weaker goal: he will be satisfied with an outcome that makes formula* $q \vee r$ *true.*

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 can improve their position by unilaterally choosing to do something 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 remain* 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. An external principal observes that it is possible to obtain an outcome in which* both *players get their goal achieved. The principal provides incentives for player* 2 *so that he will choose the more desirable outcome in which both players get their goal satisfied. The incentives are in the form of taxes: the principal taxes player* 2*'s actions so that setting* $q = \top$ *is cheaper than setting* $r = \top$*. Player 2 then chooses rationally,*

and both players get their goal satisfied. This might seem harsh on player 2*, but although he pays taxes that he would not otherwise pay, 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.*

Before proceeding, we should point out the idea of using taxation schemes to manipulate systems is by no means fanciful. In the economics literature, for example, the concept of *Tobin taxes* has been proposed as a way of stabilising foreign exchange markets [29]. Proposed by Nobel Laureate James Tobin in 1978, Tobin taxes are intended to provide incentives for foreign exchange traders to avoid excessive speculative transactions, the volume of which seems to have contributed to several crises faced by the international financial markets over the past three decades:

> It is increasingly the case that vast quantities of money flow around the global economy on pure speculation, in which the herd instinct often drives speculative waves. Invariably, given the volume of speculative flows, exchange rates overshoot their natural equilibrium, intensifying the distortions created. Such currency movements are a huge destablising force, not just for individual economies but for the global economy as a whole. [27, p.756]

Tobin taxes are taxes imposed at the level of individual transactions (also called "spot" taxes). The idea of introducing such taxes is to reduce the incentives for purely speculative transactions, thereby dampening down the overall volume of transactions, and hopefully promoting stability. A key issue in Tobin taxes is what level taxes should be set. Too low and their effect is insufficient; too high and they start to seriously impinge on the economic efficiency of the market [27, p.756]. Although our present work is not concerned with Tobin taxes, there are nevertheless striking parallels between the goals of the work in this paper and the goals of Tobin taxes.

The remainder of this paper is structured as follows:

- We begin in the following section by introducing the key concepts used throughout the remainder of the paper. We start by defining our model of Boolean games. We then define taxation schemes, and using this definition, we introduce preferences, utilities, and Nash equilibria. We then introduce *cost-free games*. Cost-free games are a simplification of our Boolean game model, which play a useful role in several of our later formal results.

- In section 3, we introduce the *implementation problem*: the problem of designing a taxation scheme so that, if players then act rationally, they will choose an outcome that satisfies the principal's objective Υ. We introduce two variants of this problem: *weak* and *strong*. The weak variant requires that Υ is satisfied in *at least one* Nash equilibrium, while the strong version requires that Υ is satisfied in *all* Nash equilibria – and moreover that there is at least one Nash equilibrium. We establish the complexity of the weak implementation problem, and give a purely logical characterisation of positive instances of weak implementation (i.e., situations in which an objective Υ can be weakly implemented). Turning to strong implementation, we again characterise the computational complexity of the problem, and

give a logical characterisation of positive instances of the strong implementation problem for an important class of taxation schemes, which we refer to as $\{0, 1\}$-taxation schemes.

- In section 4, we consider *secondary* properties of taxation schemes, relating to *social welfare* and *equity*. For example, we consider the problem of implementing Υ while minimising the total tax burden on society.
- In section 5 we study related work, and we conclude in section 6 with a brief discussion and pointers to possible future work.

Although our presentation is self-contained, some understanding of propositional logic, computational complexity [25], and game theory [24] would be helpful. Table 1 provides a summary of our main notational conventions.

2 Boolean Games and Cost-Free Games

In this section, we introduce the game models that we work with throughout the remainder of this paper. We will in fact present two variations of Boolean games. The first model generalises previous models of Boolean games [15, 3, 8], in that it explicitly introduces marginal costs for each action. The second model we present is essentially the original model of Boolean games as defined by Harrenstein *et al.* [15]; there are no costs in this latter model. To distinguish the two frameworks, we will refer to the latter class of games (dichotomous preferences without costs) as *cost-free games*. Throughout the paper, we make use of classical propositional logic, and for completeness, we thus begin by recalling the technical framework of this logic.

First, 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). Where $b \in \mathbb{B}$, we denote by $\neg b$ the negated value of b (thus if $b = \top$ then $\neg b = \bot$).

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 $\varphi \equiv \psi$; thus $\varphi \equiv \psi$ means that $\models \varphi \leftrightarrow \psi$. Note that "$\equiv$" is a meta-language relation symbol, which should not be confused with the object-language bi-conditional operator "$\leftrightarrow$".

We also make some use of the framework of *Quantified Boolean Formulae* (QBFs). QBFs are a well-known extension of propositional logic, which allows for quantification over sets of Boolean variables via expressions $\exists X \cdot \varphi$ and $\forall X \cdot \varphi$, where $X \subseteq \Phi$ is a set of Boolean variables, and φ is a formula. The QBF $\exists X \cdot \varphi$ will be true if there is some assignment of values that we

Notation	Meaning
$\mathbf{2}^S$	where S is a set, denotes the powerset of S
$\mathbb{Q}$	the rational numbers
$\mathbb{Q}_{\geq}$	the set $\{x \in \mathbb{Q} \mid x \geq 0\}$
$\mathbb{B} = \{\top, \bot\}$	the Boolean (truth) values
$\Phi = \{p, q, \ldots\}$	(finite set of) Boolean variables
$\wedge, \vee, \neg, \rightarrow, \leftrightarrow$	classical connectives (and, or, not, implies, iff)
$\mathcal{L}$	the set of propositional logic formulae over Φ
φ, ψ, χ	propositional logic formulae (i.e., elements of $\mathcal{L}$)
$v : \Phi \rightarrow \mathbb{B}$	a valuation function
$\mathcal{V}$	the set of all valuation functions
$v \models \varphi$	formula φ is satisfied by valuation v
$N = \{1, \ldots, n\}$	the players of a game (agents)
γ_i	the goal of player $i \in N$: $\gamma_i \in \mathcal{L}$
Φ_i	the Boolean variables under the control of player $i \in N$
$v_i : \Phi_i \rightarrow \mathbb{B}$	a choice for player $i \in N$
$\mathcal{V}_i$	the set of all choices for player $i \in N$
$\vec{v} = (v_1, \ldots, v_n)$	an outcome (tuple of choices, one for each player)
$c : \Phi \times \mathbb{B} \rightarrow \mathbb{Q}_{\geq}$	a marginal cost function
$G = \langle N, \Phi, c, \gamma_1, \ldots, \gamma_n, \Phi_1, \ldots, \Phi_n \rangle$	a Boolean game
$\mathcal{G}$	the set of Boolean games (over N, Φ)
$\tau : \Phi \times \mathbb{B} \rightarrow \mathbb{Q}_{\geq}$	a taxation scheme
$\mathcal{T}$	the set of taxation schemes
$c_i(v_i)$	the total marginal cost to player $i \in N$ of choice $v_i \in \mathcal{V}_i$
$\tau_i(v_i)$	tax levied on player $i \in N$ if it chose $v_i \in \mathcal{V}_i$
$e_i(v_i)$	total expense (= marginal cost + tax) of choice v_i to player i
v_i^{μ}	the most expensive choice for player $i \in N$
μ_i	the cost to player i of v_i^{μ}
$u_i(\vec{v})$	utility to player i of outcome $\vec{v}$
(G, τ)	a scenario: a Boolean game together with a taxation scheme
$NE\,(G, \tau)$	the (pure) Nash equilibria of scenario (G, τ)
$NE_{\varphi}(G, \tau)$	the Nash equilibria of (G, τ) that satisfy $\varphi \in \mathcal{L}$
$H = \langle N, \Phi, \gamma_1, \ldots, \gamma_n, \Phi_1, \ldots, \Phi_n \rangle$	a cost-free game (i.e., Boolean game without cost function)
H_G	the cost-free game extracted from Boolean game G
$NE\,(H)$	the Nash equilibria of cost-free game H
Υ	the principal's objective: $\Upsilon \in \mathcal{L}$
$WI\,(G, \Upsilon)$	the set of taxation schemes weakly implementing Υ in G
$SI\,(G, \Upsilon)$	the set of taxation schemes strongly implementing Υ in G

Table 1: Key notational conventions.

can give to the variables X such that under this assignment, φ is true; the QBF $\forall X \cdot \varphi$ is true if for all assignments of values that we can give to variables X, the formula φ is true. We allow nesting of quantifiers. Here is an example of a QBF:

$$\forall p \cdot \exists q \cdot [(p \vee q) \wedge (p \vee \neg q)] \tag{1}$$

Formula (1) in fact evaluates to false. (If p is false, there is no value we can give to q that will make the body of the formula true.) Here is another QBF:

$$\exists p \cdot \forall q \cdot [(p \vee q) \wedge (\neg p \vee \neg q)] \tag{2}$$

Formula (2) is false: whatever value is assigned to p, there is a value that can then be assigned to q so as to falsify the overall formula.

We note that the term "QBF" is often understood as referring only to formulae in prenex normal form, i.e., formula of the form

$$Q_1 x_1 \cdot Q_2 x_2 \cdots Q_k x_k \cdot \varphi(x_1, \ldots, x_k)$$

where each Q_i is a quantifier and $\varphi(x_1, \ldots, x_k)$ is a propositional formula (containing no quantifiers) over variables $x_1, \ldots, x_k$. In this paper, we will *not* require that QBF are in prenex normal form.

2.1 Boolean Games

All the games we consider are populated by a set $N = \{1, \ldots, n\}$ of *agents* – the players of the game. Each agent is assumed to have a *goal*, characterised by a satisfiable $\mathcal{L}$-formula: we write γ_i to denote the goal of agent $i \in N$. The primary aim of a player in a Boolean game is to ensure that their goal is satisfied. Each agent $i \in N$ *controls* a (possibly empty) subset Φ_i of the overall set of Boolean variables (cf. [30]). 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 $i \neq j$. (Although we do not consider it here, we note that it might be interesting in future work to investigate the possibility of variables being *jointly* controlled by agents, cf. [11]).

Where $i \in N$, a *choice* for agent i is defined by a function $v_i : \Phi_i \to \mathbb{B}$, 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. Thus $\mathcal{V}_i$ defines the *actions* or *strategies* available to agent i.

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. We will denote outcomes by $\vec{v}, \vec{v}_1$, etc. where we do not need to identify or work with their individual components.

Clearly, every outcome $\vec{v}$ uniquely defines a valuation, and we will often think of outcomes as valuations, for example writing $\vec{v} \models \varphi$ to mean that the valuation defined by the outcome $\vec{v}$ satisfies formula $\varphi \in \mathcal{L}$. Let $\varphi_{\vec{v}}$ denote the propositional logic formula that uniquely characterises

the outcome $\vec{v}$:

$$\varphi_{\vec{v}} = \left(\bigwedge_{\substack{p \in \Phi \\ \vec{v} \models p}} p \right) \wedge \left(\bigwedge_{\substack{q \in \Phi \\ \vec{v} \not\models q}} \neg q \right)$$

Let $\mathit{fulf}(\vec{v})$ denote the set of agents who have their goal fulfilled by outcome $\vec{v}$:

$$\mathit{fulf}(\vec{v}) = \{i \in N \mid \vec{v} \models \gamma_i\}.$$

Intuitively, the actions available to agents correspond to setting variables true or false. We assume that these actions have *marginal costs*, defined by a *cost function*:

$$c : \Phi \times \mathbb{B} \to \mathbb{Q}_{\geq}.$$

Thus $c(p, b)$ is the marginal cost of assigning variable $p \in \Phi$ the value $b \in \mathbb{B}$. Notice that costs are assumed to be positive rational numbers. We use rationals rather than real numbers $\mathbb{R}$ simply to avoid the tangential issues of representations for real numbers.

We will say a cost function is *uniform* if it assigns the same cost to all actions, i.e., if $\exists x \in \mathbb{Q}_{\geq}$ such that $\forall p \in \Phi, \forall b \in \mathbb{B}$ we have $c(p, b) = x$. We let c_0 denote the *zero cost function*, i.e., the uniform function that assigns zero cost to all assignments.

The 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 [15, 3]; they were introduced by Dunne *et al.* [8], where it was assumed that only the action of setting a variable to $\top$ would incur a cost (see also [26]). In fact, as we shall see later, costs are, in a limited technical sense, not required in our framework: we can capture the key strategic issues at stake without them. 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.

It may not be obvious at this point why we use costs in our model, so let us pause to reflect on this. Without costs, the only consideration a player has is whether their goal is satisfied or not. Now, suppose a particular player i is in a situation where it believes it has no possibility to get its goal achieved. What should it choose to do? If there are no other factors involved in making this decision other than whether the player's goal is achieved, then it does not matter: any choice for i will be considered equally good. But this does not seem a very realistic reflection of real-world situations. Even an individual's primary goal is not achievable, they will still typically have preferences over their choices; one very natural preference is to minimise costs. By introducing costs in the way that we do here, we can capture such preferences.

As an aside, note that the interpretation of costs that we suggest above (i.e., cost is the marginal cost of an action) is by no means the only interpretation we can give to costs. For example, we may instead interpret costs as being the value required to *maintain* a variable at some value. However, such interpretations are modelling decisions, and we will not comment on them further here.

Collecting these components together, a *Boolean game*, G, is given by a structure:

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

where:

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

We will say a game is *cost free* if it has cost function c_0.

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; their goals in turn may be dependent on the variables Φ_i, and so on. For the moment, we will postpone the formal definition of the utility functions and preferences associated with our games. However, intuitively, the idea is that if an agent has multiple ways of bringing about its goal, then it will prefer to choose one that minimises costs; and if an agent cannot get its goal fulfilled, then it simply chooses to minimise costs. Such agents have quasi-dichotomous preferences, or in other terms, lexicographic preferences whose leading criterion is goal satisfaction and whose secondary criterion is the minimization of costs.

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, an external agent known as the *principal* is assumed to be at liberty to define a taxation scheme as it sees fit. Agents will seek to minimise their overall expense, 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. The principal is assumed to have an objective that he or she desires to be satisfied. We represent this objective as a propositional formula Υ. If the principal designs the taxation scheme correctly, then agents will be incentivised to choose an outcome $\vec{v}$ so as to satisfy Υ (i.e., so that $\vec{v} \models \Upsilon$).

How exactly should we model taxation schemes? One very general approach would be to levy taxes on the basis of *outcomes*. We could model such taxes by a function

$$\tau : N \times \mathcal{V}_1 \times \cdots \times \mathcal{V}_n \to \mathbb{Q}_{\geq}$$

with the intended interpretation that $\tau(i, \vec{v})$ is the amount of tax that would be imposed on agent i if the outcome $\vec{v}$ was chosen. However, for the purposes of the present paper, we choose a simpler, *additive* model of taxes, which is action based rather than outcome based. The idea is that taxes are levied on individual actions, and the total tax imposed on an agent i is the sum of

the taxes on individual choices (assignments of truth or falsity to a variable) made in the outcome v_i chosen by i.

Formally, we therefore model a taxation scheme as a function

$$\tau : \Phi \times \mathbb{B} \to \mathbb{Q}_{\geq}$$

where the intended interpretation is that $\tau(p, b)$ is the tax that would be imposed 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 $\sum_{p \in \Phi_i} \tau(p, v_i(p))$.

To simplify some issues relating to computational complexity we make the additional innocuous assumption the values of taxation schemes can be represented with a space that is bounded by a polynomial in the size of the game (that is, the number of bits needed to represent the control assignment function, the goals, and the the cost function).

We let τ_0 denote the taxation scheme that applies no taxes to any choice, i.e., $\forall x \in \Phi$ and $b \in \mathbb{B}$, $\tau_0(x, b) = 0$.

We will say τ is a $\{0, 1\}$-*taxation scheme* if for all $p \in \Phi$ and $b \in \mathbb{B}$, we have:

1. $\tau(p, b) \in \{0, 1\}$, and
2. $\tau(p, b) = 1 - \tau(p, \neg b)$.

Notice that if τ is $\{0, 1\}$, then there is exactly one choice for every player that minimises taxes for that player (i.e., the choice where for all variables, the associated tax is 0).

Let $\mathcal{T}$ denote the set of taxation schemes (over an implicitly assumed game G).

Scenarios

We refer to pairs (G, τ) as *scenarios*: a scenario is a Boolean game modified by a taxation scheme. We say a scenario (G, τ) is *uniform* if $\exists x \in \mathbb{Q}_{\geq}$ such that $\forall p \in \Phi, \forall b \in \mathbb{B}$, $c(p, b) + \tau(p, b) = x$.

Note that a scenario can be thought of as inducing a new game, in which taxes are incorporated into the cost function. Formally, let $G = \langle N, \Phi, c, \gamma_1, \ldots, \gamma_n, \Phi_1, \ldots, \Phi_n \rangle$ be a Boolean game, and τ be a taxation scheme. Then we can define a new game G^τ with cost function c^τ defined by $c^\tau(p, b) = c(p, b) + \tau(p, b)$, and all other components of G^τ as in G. In this sense, scenarios as pairs (G, τ) are not technically required in the remainder of the paper. Nevertheless, we find it useful to deal with scenarios, as scenarios consider costs (inherent in a game) and taxes (imposed by the principal) as separate entities.

Utilities and Preferences

We will now introduce our formal model of preferences and utilities. What this model aims to capture is that taxation schemes must preserve the quasi-dichotomous nature of preferences:

- 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 outcomes that *do not* satisfy its goal, an agent prefers to minimise total expense.

Examples of such scenarios where agents have dichotomous or pseudo-dichotomous preferences can be found in several previous papers about Boolean games. See for instance Example 2 (kidney exchange) in [4]. Likewise, examples of scenarios in robot domains where each robot has a single goal (rescue someone, inspect a location and take pictures, take an object from a location and bring it back) that he should try to reach while minimizing its cost correspond to pseudo-dichtomous preferences.

Notice that taxes play a role in determining an agent's preference relation, but they are not the only factor affecting preferences. An agent's primary concern is to bring about the satisfaction of its goal; expense is a secondary concern. Thus, while taxation schemes can influence the decision making of rational agents, they cannot, ultimately, change the goals of an agent. So, if an agent has a chance to achieve its goal, it will take it, no matter what the taxation incentives are to do otherwise.

Thus the utility an agent i obtains from an outcome $\vec{v}$ is dependent on three things:

- whether the agent's goal γ_i is satisfied by $\vec{v}$ or not;
- the taxation function τ; and
- the cost function c.

The cost function and the goals of an agent are given as part of a game, but of course taxation functions are not. It is important to note that the computation of a player's utility cannot be done without a taxation scheme: the utility a player obtains is with respect to a scenario (G, τ) and an outcome $\vec{v}$ for this scenario. Sometimes, where we want to consider what might happen without a taxation scheme, we will use the zero-cost taxation scheme τ_0 to form a scenario.

To formally define preferences and utilities, we need some more notation. First, with a slight abuse of notation, we extend cost and taxation functions to players and choices as follows:

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

So, $c_i(v_i)$ is the total marginal cost to player $i \in N$ of making choice v_i, and $\tau_i(v_i)$ represents the total tax that would be levied on player i if this player made choice v_i.

We denote the total expense (= marginal cost + tax) of a choice v_i for player i by $e_i(v_i)$[1]:

$$e_i(v_i) = c_i(v_i) + \tau_i(v_i).$$

[1]This is a slight (but harmless) abuse of notation: since expenses depend on the taxation mechanism, they would be more properly denoted by $e_i(v_i, \tau)$. The taxation scheme τ will always be clear from context when we write $e_i(v_i)$.

Let v_i^{μ} denote the most expensive possible course of action for agent i:

$$v_i^{\mu} \in \arg\max_{v_i \in \mathcal{V}_i} e_i(v_i).$$

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

$$\mu_i = e_i(v_i^{\mu})$$

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

$$u_i(v_1, \ldots, v_i, \ldots, v_n) = \begin{cases} 1 + \mu_i - e_i(v_i) & \text{if } (v_1, \ldots, v_i, \ldots, v_n) \models \gamma_i \\ -e_i(v_i) & \text{otherwise.} \end{cases} \quad (3)$$

It should be clear that utility for agent i will range from $1 + \mu_i$ (the best outcome for i, where it gets its goal fulfilled by performing actions that have no tax or other cost) down to $-\mu_i$ (where i does not get its goal fulfilled but makes its most expensive choice). An agent will always get utility ≥ 1 if its goal is satisfied in an outcome, and utility ≤ 0 if its goal is not satisfied.

Utility functions are defined for the sake of convenience: they provide a convenient numeric representation of preference relations. But note that our framework is purely ordinal: any order-preserving transformation of our utility functions would work equally well, and the quantitative nature of utility will not play any role in our settings (in particular, utility is not transferable).

Nash Equilibrium

Given this formal definition of utility, we can define solution concepts in the standard game-theoretic way (see, e.g., [24]). In this paper, we focus on the concept of (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 $(v_1, \ldots, v_i, \ldots, v_n)$ with respect to a scenario (G, τ) is a Nash equilibrium if for each agent $i \in N$, there is no alternative choice $v_i' \in \mathcal{V}_i$ for i such that $u_i(v_1, \ldots, v_i', \ldots, v_n) > u_i(v_1, \ldots, v_i, \ldots, v_n)$. We will let $NE(G, \tau)$ denote the set of all Nash equilibria of the scenario (G, τ). If we have some outcome $\vec{v}$ such that $\vec{v} \notin NE(G, \tau)$ then we say that the outcome $\vec{v}$ is *unstable*. An outcome is unstable if some player can benefit by unilaterally deviating from the outcome: the benefit may be either in the form of reduced expense, or in the form of getting a goal fulfilled where it was not otherwise. Where (G, τ) is a scenario and $\varphi \in \mathcal{L}$, then we will write $NE_{\varphi}(G, \tau)$ to denote the Nash equilibria of (G, τ) that satisfy φ:

$$NE_{\varphi}(G, \tau) = \{\vec{v} \in NE(G, \tau) \mid \vec{v} \models \varphi\}.$$

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.

Observation 1 *If* $(v_1^*, \ldots, v_i^*, \ldots, v_n^*) \in NE(G, \tau)$ *is such that* $i \notin fulf(v_1^*, \ldots, v_i^*, \ldots, v_n^*)$ *then:*

$$v_i^* \in \arg\min_{v_i \in \mathcal{V}_i} e_i(v_i)$$

Proof: Agent i has no choice v_i' s.t. $(v_1^*, \ldots, v_i', \ldots, v_n^*) \models \gamma_i$, else $u_i(v_1^*, \ldots, v_i', \ldots, v_n^*) > u_i(v_1^*, \ldots, v_i^*, \ldots, v_n^*)$, in which case $(v_1^*, \ldots, v_i^*, \ldots, v_n^*) \notin NE(G, \tau)$. So, the only way i could profitably deviate would be by making an alternative choice v_i' that reduced expense compared to v_i^*. But by definition, v_i^* minimises i's expense. ∎

The following result establishes a sufficient condition for the instability of outcomes; we will use this result later.

Observation 2 *Suppose $(v_1, \ldots, v_i, \ldots, v_n)$ is an outcome for a scenario (G, τ). Then if there is some player i such that $i \notin fulf(v_1, \ldots, v_i, \ldots, v_n)$ and $\exists v_i' \in \mathcal{V}_i$ such that $i \in fulf(v_1, \ldots, v_i', \ldots, v_n)$, then $(v_1, \ldots, v_i, \ldots, v_n) \notin NE(G, \tau)$.*

Proof: From (3), we have $u_i(v_1, \ldots, v_i, \ldots, v_n) \leq 0$ while $u_i(v_1, \ldots, v_i', \ldots, v_n) \geq 1$, so i can benefit by unilaterally deviating from $(v_1, \ldots, v_i, \ldots, v_n)$. ∎

Notice that this result establishes the instability of an outcome *independently* of the cost function c and taxation scheme τ of the scenario. These play no part in the choice of the player in this setting: the player will *always* prefer to unilaterally deviate from an outcome if such deviation results in a goal being fulfilled where it would not otherwise be fulfilled.

The following is an natural decision problem to consider:

NASH VERIFICATION:
Instance: Boolean game G, taxation scheme τ, and outcome $\vec{v}$.
Question: Is it the case that $\vec{v} \in NE(G, \tau)$?

Proposition 1 NASH VERIFICATION *is co-*NP*-complete.*

Proof: Membership is immediate. Hardness is shown by reducing the complement problem (checking that an outcome is not a Nash equilibrium) from SAT. This is the problem of determining whether a given propositional logic formula φ is satisfiable, i.e., true under at least one valuation. Given an instance φ of SAT over variables $x_1, \ldots, x_k$, define a game G with $N = \{1\}$, $\Phi = \Phi_1 = \{x_1, \ldots, x_k, z\}$ (where z does not occur in φ), $\gamma_1 = \varphi \wedge z$, and v the valuation that makes all variables false, i.e., $v \models \neg z \wedge \neg x_1 \wedge \cdots \wedge \neg x_k$. Define all costs and taxes to be 0. Player 1 then has a beneficial deviation from v iff φ is satisfiable (in which case the beneficial deviation is to fulfil its goal, making $\varphi \wedge z$ true). ∎

Next, note that while being able to model costs in games explicitly is attractive from a modelling perspective, it is, in a sense, unnecessary from a purely technical point of view: we can always design a taxation scheme that simulates the costs and thus gives rise to the same set of Nash equilibria.

Proposition 2 *Let G be a game with cost function c and let τ be a taxation scheme for G. Then there exists a taxation scheme τ' such that $NE(G, \tau) = NE(G', \tau')$ for the game G' we obtain by replacing c with c_0 in G.*

Proof: Let $\tau'(p, b) = \tau(p, b) + c(p, b)$ for all $p \in \Phi$ and all $b \in \mathbb{B}$. Then the utility functions for (G', τ') are identical to those for (G, τ), and thus the Nash equilibria must coincide as well. ∎

Moreover, we can show that, for the analysis of Nash equilibria, it suffices to consider taxation schemes that only impose taxes on making a variable *true* (rather than *false*). Call a taxation scheme τ *positive* if $\tau(p, \bot) = 0$ for all $p \in \Phi$. Now consider two zero cost games G and G'[2]. We call G' a *variant* of G if G' is the same as G, except that for some $p \in \Phi$ all occurrences of p in the agents' goals γ_i have been replaced by $\neg p$ (but Υ has not been changed).

Proposition 3 *Let G be a zero cost game and let τ be a taxation scheme for that game. Then there exists a variant G' of G and a positive taxation scheme τ' such that $NE(G, \tau) = NE(G', \tau')$.*

Proof: We have to define G' and τ' with respect to each $p \in \Phi$. For all variables p we will have $\tau'(p, \bot) = 0$ (as τ' should be positive). So we have to define the values $\tau'(p, \top)$ and we have to specify whether p should occur in the goal formulas in G' as in G, or whether p should get flipped (i.e., whether it should get rewritten as $\neg p$).

1. If $\tau(p, \top) = \tau(p, \bot)$, then we set $\tau'(p, \top) = 0$ and we leave p untouched in the game.
2. If $\tau(p, \top) > \tau(p, \bot)$, then we set $\tau'(p, \top) = \tau(p, \top) - \tau(p, \bot)$ and we again leave p untouched in the game.
3. If $\tau(p, \top) < \tau(p, \bot)$, then we set $\tau'(p, \top) = \tau(p, \bot) - \tau(p, \top)$ and we flip p in the game.

The crucial feature of this construction is that the difference in tax between making p *true* or *false* remains $|\tau(p, \top) - \tau(p, \bot)|$ in the new game, and the new taxation scheme still "pushes in the same direction" as before. Therefore, the utility functions for (G', τ') are identical to those for (G, τ), and thus the Nash equilibria must coincide as well. ∎

2.2 Cost-free Games

The model of Boolean games presented above represents the basic model of games that we will work with in this paper. However, we occasionally find it useful to consider a special class of Boolean games which we refer to as *cost-free games* [14]. A cost-free game is a Boolean game with dichotomous preferences, i.e., with no costs or taxes: a player in a cost-free game is *only* concerned with whether their goal is satisfied or not. Formally, a cost-free game, H, is given by a structure

$$H = \langle N, \Phi, \gamma_1, \ldots, \gamma_n, \Phi_1, \ldots, \Phi_n \rangle$$

with components as in Boolean games, above. Thus, a cost-free game is a Boolean game with no marginal cost function c. Outcomes are defined for cost-free games in exactly the same way as for Boolean games, i.e., an outcome is a tuple of choices, one for each player in the game:

[2]This restriction to games with zero cost is not required, but it does simplify exposition; and we have just seen that for the analysis of Nash equilibria it suffices to consider zero cost games.

$\vec{v} \in \mathcal{V} \times \cdots \times \mathcal{V}_n$. Because no costs or taxes are involved, the definition of utility used for cost-free games is much simpler than that used for our more general model of Boolean games. Formally, (and with a small abuse of notation), the utility for player $i \in N$ of an outcome $\vec{v}$ is denoted by $u_i(\vec{v})$, and is defined as follows:

$$u_i(\vec{v}) = \begin{cases} 1 & \text{if } \vec{v} \models \gamma_i \\ 0 & \text{otherwise.} \end{cases}$$

With a further abuse of notation, we will write $NE(H)$ to denote the set of Nash equilibria of the cost-free game H.

We will denote by H_G the cost-free game obtained from Boolean game G by simply removing the cost function. The following results relate the Nash equilibria of Boolean games and cost-free games, and will be useful later.

Proposition 4 *Let G be a Boolean game, and let $\tau \in \mathcal{T}$ be such that scenario (G, τ) is uniform. Then $NE(G, \tau) = NE(H_G)$. It follows that for all Boolean games G, there exists a taxation scheme $\tau \in \mathcal{T}$ such that $NE(G, \tau) = NE(H_G)$.*

Proof: For each agent i, let u_i and u_i' the utility functions induced respectively by G and (G, τ). Because (G, τ) is uniform, for every i there is a x_i such that $c(v_i) + \tau(v_i) = x_i$ for all $v_i \in \mathcal{V}_i$. Let v be a valuation. If $v \models \gamma_i$ then $u_i(v) = u_i'(v) = 1$, and if $v \models \neg\gamma_i$ then $u_i(v) = -x_i$ whereas $u_i'(v) = 0$. Therefore, for each agent i, u_i and u_i' are comonotonic, from which we conclude that the Nash equilibria of G and (G, τ) coincide. The second item follows immediately. ∎

Proposition 5 *For all Boolean games G, and for all $\tau \in \mathcal{T}$, we have $NE(G, \tau) \subseteq NE(H_G)$. Moreover, there exist Boolean games G and taxation schemes τ for which the inclusion is strict.*

Proof: For the first part, let G, τ, and H_G be as in the statement of the proposition, and suppose $\vec{v} \in NE(G, \tau)$. Then no player has a beneficial deviation from $\vec{v}$, i.e., no player can unilaterally make an alternative choice that would either reduce costs or lead to the achievement of its goal. But since no player can unilaterally deviate to fulfil its goal, then $\vec{v} \in NE(H_G)$. The second part is immediate: by introducing a taxation scheme, we can eliminate some Nash equilibria. ∎

3 The Implementation Problem

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. The principal will do this for a reason: because it wants to provide incentives for the players 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*. The implementation problem inherits concepts from the theory of Nash implementation in mechanism design [18], although our use of Boolean games, taxation schemes, and propositional formulae to represent objectives is quite different.

We define two variants of the implementation problem, as follows:

- In the WEAK IMPLEMENTATION problem, we are given a Boolean game G and an objective Υ, and we are asked whether there exists any taxation scheme τ such that the scenario (G, τ) has at least one Nash equilibrium satisfying Υ.
- In the STRONG IMPLEMENTATION problem, we are given a Boolean game G and an objective Υ, and we are asked whether there exists any taxation scheme τ such that:
  1. the scenario (G, τ) has at least one Nash equilibrium; and
  2. all Nash equilibria of (G, τ) satisfy Υ.

It is worth noting that the design of a taxation scheme is done by the principal with full knowledge of the game G. In particular, we assume here that the principal knows the goals and possible actions of the players of the game G.

In what follows, we formally define these two variants of the implementation problem, and explore some issues surrounding them.

3.1 Weak Implementation

Let $WI(G, \Upsilon)$ denote the set of taxation schemes τ over G such that in the resulting scenario (G, τ), the propositional objective formula Υ is satisfied in at least one Nash equilibrium outcome:

$$WI(G, \Upsilon) = \{\tau \in \mathcal{T} \mid NE_{\Upsilon}(G, \tau) \neq \emptyset\}.$$

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 $WI(G, \Upsilon) \neq \emptyset$?

If $\langle G, \Upsilon \rangle$ is a positive instance of the WEAK IMPLEMENTATION problem, then we say that Υ *can be weakly implemented in* G.

Let us see an example.

Example 2 *Define a game G as follows: $N = \{1, 2\}$, $\Phi = \{p_1, p_2\}$, $\Phi_i = \{p_i\}$, $\gamma_1 = p_1$, $\gamma_2 = \neg p_1 \wedge \neg p_2$, $c(p_1, b) = 0$ 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 fulfilled, 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 fulfilled, 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). Let k be an astronomically large value. Then define a taxation scheme τ such that:
>
> $$\tau(p, b) = \begin{cases} 0 & \text{if } b = v(p) \\ k & \text{otherwise.} \end{cases}$$

Thus, the idea is simply to make all choices other than selecting the outcome corresponding to v, which 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 its goal over an outcome that does: see Observation 2. The following example illustrates the fact that satisfiable objectives cannot always be implemented.

Example 3 *Define a game G as follows: $N = \{1, 2\}$, $\Phi = \Phi_1 = \{p\}$, $\Phi_2 = \emptyset$, $\gamma_1 = p$, $\gamma_2 = \top$, and $c(p, b) = 0$ for all $b \in \mathbb{B}$. Let $\Upsilon = \neg p$. Suppose there is a taxation scheme τ such that $\exists \vec{v} \in NE(G, \tau)$ and $\vec{v} \models \Upsilon$, i.e., $\vec{v} \models \neg p$. So $\vec{v} \not\models \gamma_1$ and thus $u_1(\vec{v}) = -e_1(v_1) = 0$. But suppose 1 chose the valuation v_1 with $v_1(p) = \top$. Clearly this choice would satisfy its goal, and would hence represent a profitable deviation from $\vec{v}$ for 1, so $\vec{v} \notin NE(G, \tau)$; contradiction.*

So, satisfiability is not a sufficient condition for implementability. But what about tautologies, i.e., objectives Υ such that $\Upsilon \equiv \top$? We might be tempted to assume that tautologies are trivially implementable. This is not in fact the case, however, because for weak implementation, we require that the objective Υ is satisfied *in a Nash equilibrium outcome*, and it may be that $NE(G, \tau) = \emptyset$ for all taxation schemes τ. Consider the following example.

Example 4 *Define a game G with $N = \{1, 2\}$, $\Phi = \{p, q\}$, $\Phi_1 = \{p\}$, $\Phi_2 = \{q\}$, $\gamma_1 = (p \leftrightarrow q)$, $\gamma_2 = \neg(p \leftrightarrow q)$, and c assigns zero cost to all actions. We have $NE(G, \tau) = \emptyset$ for all taxation schemes τ. This is because for every outcome $(v_1, v_2) \in \mathcal{V}_1 \times \mathcal{V}_2$ some unsuccessful player can deviate to get their goal satisfied. Hence by Observation 2, every outcome is unstable irrespective of the taxation scheme chosen. For example, in the outcome (v_1, v_2) in which $v_1(p) = \top$ and $v_2(q) = \top$, agent 2 would benefit by changing its valuation to $v_2'(q) = \bot$. (Alert readers will note that this is a Boolean games version of the well-known game of matching pennies.)*

Tautologous objectives (i.e., objectives Υ such that $\Upsilon \equiv \top$) might appear to be of little interest, but this is not the case. For suppose we have a game G such that $NE(G, \tau_0) = \emptyset$. Then, in its unmodified condition, before we impose any taxation scheme, 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 5 *Let $N = \{1, 2, 3\}$, with $\Phi = \{p, q, r\}$, $\Phi_1 = \{p\}$, $\Phi_2 = \{q\}$, $\Phi_3 = \{r\}$, $\gamma_1 = \top$, $\gamma_2 = (q \wedge \neg p) \vee (q \leftrightarrow r)$, $\gamma_3 = (r \wedge \neg p) \vee \neg(q \leftrightarrow r)$, $c(p, \top) = 0$, $c(p, \bot) = 1$, 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.*

Now, we know from the work of Bonzon *et al.* [3] that the problem of checking for the existence of pure strategy Nash equilibria in cost-free games is Σ_2^p-complete. It turns out that the WEAK IMPLEMENTATION problem is no harder:

Proposition 6 *The* STABILISATION *problem is Σ_2^p-complete. As a consequence, the* WEAK IMPLEMENTATION *problem is also Σ_2^p-complete.*

Proof: Membership requires evaluating the following condition:

$$\exists \tau \in \mathcal{T} : \exists \vec{v} \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_n : [\vec{v} \in NE(G, \tau)].$$

Notice that checking the inner condition, $\vec{v} \in NE(G, \tau)$, is a co-NP problem (Proposition 1), and that the existential quantifiers refer to structures that are of size polynomial in the size of the input. Thus the problem is in Σ_2^p. Hardness follows straightforwardly from the fact that checking for the existence of pure strategy Nash equilibria in cost-free Boolean games is Σ_2^p-complete [3, Proposition 5]. ∎

Characterising Weak Implementability

A very natural question to ask is whether we can give a *logical characterisation* of weak implementability. That is, whether we can obtain a purely logical condition Ψ on games G and objectives Υ (i.e., a condition that depends only on the logical structure of the game, and is independent of taxation schemes and costs), such that Υ can be weakly implemented in G iff property Ψ holds of G and Υ. In this section, we answer this question in the affirmative.

As a warm-up exercise, let us obtain a *sufficient* condition for weak implementation, as follows. Suppose we can find a valuation that satisfies Υ *and* the goal of every player i: then the only reason a player would deviate from this valuation would be to reduce costs. But we can eliminate such temptations by fixing the taxation scheme to make any such deviation unprofitable. Formally, we have:

Proposition 7 *For all games G and objectives $\Upsilon \in \mathcal{L}$, if the formula Υ' is satisfiable:*

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

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

Proof: Assume $\Upsilon' = \Upsilon \wedge \bigwedge_{i \in N} \gamma_i$ is satisfiable. Let v be a valuation such that $v \models \Upsilon'$. The basic idea is to use the approach (*), described above, to build a taxation scheme ensuring that the valuation v is a rational choice. For all $i \in N$, $p \in \Phi_i$ and $b \in \mathbb{B}$, define:

$$\tau(p, b) = \begin{cases} 0 & \text{if } b = v(p) \\ 1 + k_i & \text{otherwise.} \end{cases}$$

where k_i is the cost of the choice for i that has the highest marginal cost[3]; that is:

$$k_i = \max\{c_i(v_i) \mid v_i \in \mathcal{V}_i\}.$$

Let $\vec{v}$ be the outcome corresponding to the valuation v. Obviously, $\vec{v} \models \Upsilon$. We claim that $\vec{v} \in NE(G, \tau)$. For suppose that $\vec{v}$ is not a Nash equilibrium. Then some agent i can benefit by deviating. Since by construction $\vec{v} \models \gamma_i$, then i can only benefit from a choice that would decrease its overall costs. But the construction of τ ensures that any other choice would *increase* taxes, and hence overall expense, more than any benefit gained by decreasing marginal costs. So, i cannot benefit by changing its choice, and so $\vec{v}$ is a Nash equilibrium. ∎

We can therefore describe informally our ability to influence behaviour through taxes: we can eliminate through taxation the temptation to deviate from an outcome by reducing costs, but we *cannot* eliminate the temptation to deviate to get a goal fulfilled. Our characterisation result is thus as follows:

Proposition 8 *Let G be a Boolean game and $\Upsilon \in \mathcal{L}$ be an objective. Then Υ can be weakly implemented in G iff there exists an outcome $(v_1, \ldots, v_i, \ldots, v_n) \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_i \times \cdots \times \mathcal{V}_n$ such that:*

1. *$(v_1, \ldots, v_i, \ldots, v_n) \models \Upsilon$; and*
2. *for every $i \notin fulf(v_1, \ldots, v_i, \ldots, v_n)$ there is no $v_i' \in \mathcal{V}_i$ such that $(v_1, \ldots, v_i', \ldots, v_n) \models \gamma_i$.*

Proof:

($\rightarrow$) Assume that Υ can be weakly implemented in G, and let $\tau \in \mathcal{T}$ be a taxation scheme that weakly implements Υ. By definition, $\exists (v_1, \ldots, v_i, \ldots, v_n) \in NE(G, \tau)$ such that $(v_1, \ldots, v_i, \ldots, v_n) \models \Upsilon$. Now suppose for contradiction that $\exists i \notin fulf(v_1, \ldots, v_i, \ldots, v_n)$ and $\exists v_i' \in \mathcal{V}_i$ such that $(v_1, \ldots, v_i', \ldots, v_n) \models \gamma_i$. From Observation 2, $(v_1, \ldots, v_i, \ldots, v_n) \notin NE(G, \tau)$; contradiction.

[3] Note that the value k_i here is not the same as the value μ_i defined earlier.

($\leftarrow$) Assume $\exists \vec{v} = (v_1, \ldots, v_i, \ldots, v_n)$ such that $\vec{v} \models \Upsilon$ and that $\forall i \notin \mathit{fulf}(\vec{v})$ it is not the case that $\exists v_i' \in \mathcal{V}_i$ such that $(v_1, \ldots, v_i', \ldots, v_n) \models \gamma_i$. Define a taxation scheme τ as in Proposition 7, so that for every player i, every choice other than choosing v_i is strictly more expensive than choosing v_i. Then $\vec{v} \in \mathit{NE}(G, \tau)$: indeed, assume i has an interest to deviate; since his course of action minimizes the expense (cost + tax) then the only possibility for him to improves his utility is to have $\vec{v} \models \neg\gamma_i$ and $(v_1, \ldots, v_i', \ldots, v_n) \models \gamma_i$, which by assumption is not possible.

∎

Notice that, as in Proposition 2, the condition of Proposition 8 is independent of the cost function c in a scenario. Thus, weak implementability in a Boolean game is equivalent to weak implementability in the induced cost-free game H_G: to determine whether Υ can be weakly implemented in G it suffices to determine whether it can be weakly implemented in the cost-free version of G. We obtain:

Proposition 9 *Let G be a Boolean game and let $\Upsilon \in \mathcal{L}$ be an objective. Then Υ can be weakly implemented in G iff $\exists \vec{v} \in \mathit{NE}(H_G)$ such that $\vec{v} \models \Upsilon$.*

Now, Bonzon *et al.* were able to give a logical characterisation of the Nash equilibria of (cost free) Boolean games, in terms of the truth of QBFs [6, Proposition 3]. We can adapt this result to give a QBF characterisation of weakly implementable objectives, as follows:

Proposition 10 *Let G be a Boolean game and $\Upsilon \in \mathcal{L}$ be an objective. Then Υ can be weakly implemented in G iff the following QBF is satisfiable:*

$$\Upsilon \wedge \bigwedge_{i \in N} ((\exists \Phi_i \cdot \gamma_i) \rightarrow \gamma_i$$

Proof: The claim follows from Proposition 8. The formula is a direct translation of the two conditions in Proposition 8 into a QBF. The formula is satisfiable iff there exists a valuation that satisfies Υ, and that for that same valuation, it must be the case that any agent who is left unsatisfied by this valuation cannot change her part of the valuation in such a way that the resulting valuation will satisfy her goal. ∎

As an aside, this QBF characterization of weak implementability implies that the complexity of WEAK IMPLEMENTATION can be reduced whenever the complexity of the corresponding QBF problem can be reduced. This holds in particular when all the goals γ_i are under Disjunctive Normal Form (DNF). In this case, determining whether $\exists \Phi_i \cdot \gamma_i$ can be computed in polynomial time, and in this case WEAK IMPLEMENTATION is "only" NP-complete.

Example 6 *Suppose we have a game with $N = \{1, 2, 3\}$, $\Phi_1 = \{a\}$, $\Phi_2 = \{b\}$, $\Phi_3 = \{c\}$, and*

$\gamma_1 = a \vee b$;

$\gamma_2 = a \leftrightarrow (b \leftrightarrow c)$;

$$\gamma_3 = (a \wedge \neg b \wedge \neg c) \vee (\neg a \wedge b \wedge c).$$

Now, we have:

$$\begin{array}{l} \exists \Phi_1 \cdot \gamma_1 = \exists a \cdot a \vee b \equiv \top \\ \exists \Phi_2 \cdot \gamma_2 = \exists b \cdot a \leftrightarrow (b \leftrightarrow c) \equiv \top \\ \exists \Phi_3 \cdot \gamma_3 = \exists c \cdot (a \wedge \neg b \wedge \neg c) \vee (\neg a \wedge b \wedge c)) \equiv (a \wedge \neg b) \vee (\neg a \wedge b) \end{array}$$

We invite the reader to verify that:

$$\bigwedge_i (\gamma_i \vee \neg \exists \Phi_i . \gamma_i) \equiv (a \leftrightarrow \neg b) \wedge \neg c$$

And so Υ is weakly implementable if the following formula is satisfiable:

$$\Upsilon \wedge ((a \leftrightarrow \neg b) \wedge \neg c$$

The logical characterisation of weak implementability given in Proposition 10 shows that WEAK IMPLEMENTABILITY could be solved by a QBF solver.

How Much Taxation Do We Need?

We now show that, if an objective can be implemented, then it can be implemented using only a *small* amount of taxation. We will first prove a slightly weaker result.

Proposition 11 *Let G be a game with a uniform cost function, let $\Upsilon \in \mathcal{L}$ be an objective, and let $\varepsilon \ll 1$ be any arbitrarily small positive rational number. Then Υ is weakly implementable in G iff Υ can be weakly implemented by a taxation scheme τ such that $\tau(x,b) \leq \varepsilon$ for all x, b.*

Proof:

($\rightarrow$) Assume $\tau \in WI(G, \Upsilon)$ and let $\vec{v} = (v_1, \ldots, v_n) \in NE(G, \tau)$ be such that $\vec{v} \models \Upsilon$. Now, define a taxation scheme τ^* as follows. For each $i \in N$, $p \in \Phi_i$, and $b \in \mathbb{B}$:

$$\tau^*(p,b) = \begin{cases} \varepsilon & \text{if } v_i(p) \neq b \\ 0 & \text{otherwise.} \end{cases}$$

We claim that $\tau^* \in WI(G, \Upsilon)$. For consider the outcome $(v_1, \ldots, v_n)$. Since $(v_1, \ldots, v_n) \in NE(G, \tau)$, then from Observation 2, no unsuccessful player i can deviate from v_i with an alternative choice v_i' and thereby achieve goal γ_i, so the only possible beneficial deviation for i would be to reduce costs; but the construction of τ^* ensures that such a deviation would result in greater taxes for i than choosing v_i.

($\leftarrow$) Immediate.

∎

Now, this result applies to games with a uniform cost function. But intuitively, we can *make* a game uniform by setting taxes appropriately, to "level out" taxes and costs; by then using the idea of the preceding proposition, we only then need some small additional tax to "fix" the desired outcome.

Proposition 12 *Let G be a game, let $\Upsilon \in \mathcal{L}$ be an objective, and let $\varepsilon \ll 1$ be any arbitrarily small positive number. Then Υ is weakly implementable in G iff Υ can be weakly implemented by a taxation scheme τ^* such that τ^* is bounded by the value:*

$$\varepsilon + \max\{c_i(v_i) \mid i \in N,\ v_i \in \mathcal{V}_i\}$$

Proof:

($\rightarrow$) First, define the value t^* as follows:

$$t^* = \max\{c_i(v_i) \mid i \in N,\ v_i \in \mathcal{V}_i\}$$

Next, define a taxation scheme τ_1 as follows. For all $p \in \Phi$, $b \in \mathbb{B}$:

$$\tau_1(p, b) = t^* - c(p, b).$$

It is clear from construction that the scenario (G, τ_1) is uniform. Now, let τ_2 be a taxation scheme that weakly implements Υ in G, and let $(v_1, \ldots, v_n)$ be an outcome such that $(v_1, \ldots, v_n) \in NE(G, \tau_2)$ and $(v_1, \ldots, v_n) \models \Upsilon$. The outcome $(v_1, \ldots, v_n)$ is the one we want to "fix", by making all other choices more expensive. To do this, we define a third taxation scheme τ_3 as follows. For all $i \in N$, $p \in \Phi_i$, and $b \in \mathbb{B}$:

$$\tau_3(p, b) = \begin{cases} \varepsilon + \tau_1(p, b) & \text{if } v_i(p) \neq b \\ \tau_1(p, b) & \text{otherwise.} \end{cases}$$

We claim that $(v_1, \ldots, v_n) \in NE(G, \tau_3)$, and hence that $\tau_3 \in WI(G, \Upsilon)$. For suppose that $(v_1, \ldots, v_n) \notin NE(G, \tau_3)$. Then some player can benefit by making a different choice: either to get its goal fulfilled, or to reduce costs. No unsuccessful player can deviate and get his goal fulfilled, for otherwise from Observation 2 we would have $(v_1, \ldots, v_n) \notin NE(G, \tau_2)$. But equally, by construction, no player can reduce costs from deviation.

($\leftarrow$) Immediate.

∎

Now, at this point let us return to the assumption we made earlier with respect to the space requirements for taxes: that we are restricting our attention to taxation schemes requiring space polynomial in the size of the game. The preceding proposition shows that if we can weakly implement Υ, then we can weakly implement it with only "small" taxes. However, this does not in itself imply that the *space requirement* for the "small" taxation scheme is also small: we could conceivably have some small tax value that required a large amount of space to represent it! Nevertheless, it should be clear that if we restrict our attention to games with costs that are given by natural numbers, then essentially the same construction can be used to build taxation schemes that *do* have small space requirements.

Dependence and Weak Implementation

Example 1 refers to an implicit notion of *task exchanges* between agents, as in [2, 13], or more generally of *dependence* between agents. In Example 1, for instance, player 1 depends on player 2 for the achievement of her goal. A generalization of this example leads to make the notion of dependence more formal and to exploit it for the weak implementation problem in the following way: if B is a subset of agents; if the agents in B, whatever their internal interdependencies, do not rely on agents outside of B to have their goals achieved, and if their goals form a maximal consistent subset of the set of all goals, then the conjunction of their goals is weakly implementable.

Formally, this calls for a notion of dependence. We reuse the notion of dependence from [5] and [26]: player i depends on player j (for Boolean game G), denoted by iD_Gj, if some variable controlled by j is relevant to γ_i. As in [5] we define a stable set for G as a set $B \subseteq N$ such that for all $i \in B$, every relevant variable for γ_i is controlled by an agent in B. Then we have the following result:

Proposition 13 *Let G be a Boolean game with costs. If $B \subseteq N$ is stable for G and $\gamma_B = \{\gamma_i \mid i \in B\}$ is maximal consistent in $\{\gamma_i \mid i \in N\}$ then γ_B is weakly implementable in G.*

Proof: Assume (1) $B \subseteq N$ is stable for G and (2) $\gamma_B = \{\gamma_i \mid i \in B\}$ is maximal consistent in $\{\gamma_i \mid i \in N\}$. Let $v \models \gamma_B$ (such a v exists because γ_B is consistent). Consider the following taxation function: for all $i \in B$ and each variable p controlled by i: if $v_i \models p$ then $\tau(p) = 0$ and $\tau(\neg p) = \max(0, c(p) - c(\neg p)) + 1$; if $v_i \models \neg p$ then $\tau(\neg p) = 0$ and $\tau(p) = \max(0, c(\neg p) - c(p)) + 1$; all other taxes are 0. We claim that v is a Nash equilibrium of $G + \tau$. Obviously, no agent in B has an incentive to deviate from v since they have their goal satisfied and pay a zero tax. Suppose an agent $i \notin B$ has an interest to deviate from v by playing $v'_i \neq v_i$; let $v' = (v_{-i}, v'_i)$. Since i pays a zero tax in v, the only reason for him to deviate is to have his goal satisfied in v'. Now, since the goals of the agents in B depend only on the variables they control, $v \models \gamma_B$ implies $v' \models \gamma_B$; therefore, $v \models \gamma_B \wedge \gamma_i$, which contradicts the assumption that γ_B is maximal consistent in $\{\gamma_i \mid i \in N\}$. Therefore v is a Nash equilibrium of $G + \tau$ and since $s \models \gamma_B$, we have shown that γ_B is strongly implementable in G. ∎

As a consequence, if γ_N is consistent then it is weakly implementable in G – which we knew already from Proposition 9, letting $\Upsilon = \gamma_N$. More generally, when $B \neq N$, B being stable implies that B is powerful enough to have their joint goals achieved – possibly via exchange of favours as in [2], but not necessarily.

Note that we need both conditions (1) and (2) for the result to hold. Consider the following Boolean G game where (1) is not satisfied: $N = \{1, 2\}$, 1 (resp. 2) controls x_1 (resp. x_2), $\gamma_1 = x_2$, $\gamma_2 = \neg x_2$; no costs. $\{\gamma_1\}$ is maximal consistent but not stable; γ_1 is not weakly implementable in G. Consider now the following Boolean game G': $N = \{1, 2, 3, 4\}$, i controls x_i for $i = 1, 2, 3, 4$, $\gamma_1 = x_2$, $\gamma_2 = x_1$, $\gamma_3 = x_3 \leftrightarrow x_4$, $\gamma_4 = x_3 \leftrightarrow \neg x_4$; no costs. The set $\{1, 2\}$ is stable and $\gamma_{\{1,2\}}$ is consistent, but not maximal consistent, since $\{\gamma_1, \gamma_2, \gamma_3\}$ is consistent; G' is not stabilisable, *a fortiori* $\gamma_1 \wedge \gamma_2$ is not weakly implementable in G'.

3.2 Strong Implementation

The fact that $WI(G, \Upsilon) \neq \emptyset$ is good news for the principal – it says that the principal can impose a taxation scheme such that *at least one* rational (NE) outcome of the resulting scenario satisfies Υ. However, it could be that there are many Nash equilibria, and only one of them satisfies Υ. If we design taxation schemes appropriately, then we can "steer" players towards outcomes that satisfy Υ in equilibrium, by making other choices too expensive. But nevertheless, the presence of other Nash equilibria, which potentially do not satisfy Υ, is clearly undesirable. This motivates us to consider the *strong implementation* problem. Strong implementation is similar to the notion of Nash implementation in the mechanism design literature [18]. With strong implementation, we require not only that the objective Υ is satisfied in some Nash equilibrium outcome, but that it is also satisfied in *all* equilibrium outcomes.

Let $SI(G, \Upsilon)$ denote the set of taxation schemes τ over G such that:

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

Formally:

$$SI(G, \Upsilon) = \{\tau \in \mathcal{T} \mid NE(G, \tau) \neq \emptyset \ \& \ \forall \vec{v} \in NE(G, \tau) : \vec{v} \models \Upsilon\}.$$

This gives us the following decision problem:

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

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

Proposition 14 STRONG IMPLEMENTATION *is* Σ_2^p*-complete.*

Proof: Observe that the problem involves evaluating the following condition: is it the case that $\exists \tau \in \mathcal{T} : NE(G, \tau) \neq \emptyset$ and $\forall \vec{v} \in NE(G, \tau)$ we have $\vec{v} \models \Upsilon$? Expanding out and re-arranging, it can be seen that this is equivalent to asking whether $\exists \tau \in \mathcal{T}, \exists \vec{v} \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_n, \forall \vec{v}' \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_n$, we have $\vec{v} \in NE(G, \tau)$ and if $\vec{v}' \in NE(G, \tau)$ we have $\vec{v}' \models \Upsilon$. Clearly this is a Σ_2^p predicate. For hardness, we can reduce the STABILISATION problem as in Proposition 6. ∎

How are $WI(G, \Upsilon)$ and $SI(G, \Upsilon)$ related? It turns out that weak and strong implementation are indeed different. It is obvious that for all games G and objectives Υ we have $SI(G, \Upsilon) \subseteq WI(G, \Upsilon)$, but as the following example illustrates, there exist games G and objectives Υ such that $WI(G, \Upsilon) \not\subseteq SI(G, \Upsilon)$.

Example 7 *We give an example of a game and an objective such that the objective can be weakly, but not strongly implemented. Let* $N = \{1, 2\}$*, with* $\Phi = \{p, q\}$*,* $\Phi_1 = \{p\}, \Phi_2 = \{q\}$*,* $\gamma_1 = \gamma_2 = (p \leftrightarrow q)$*, with cost function* c_0*. Finally, let* $\Upsilon = p \wedge q$*. Now, the taxation function* τ_0 *with zero taxes will weakly implement* Υ*: there will be two Nash equilibria, one satisfying* $p \wedge q$ *and*

the other satisfying $\neg(p \vee q)$. However, $\Upsilon = p \wedge q$ cannot be strongly implemented, because the outcome satisfying $\neg(p \vee q)$ will be a Nash equilibrium for all taxation schemes τ. To see this, observe that for it not to be a Nash equilibrium, one agent would benefit by deviating; but by definition of the utility functions, such a deviation would involve an agent moving from positive to negative utility.

Characterising Strong Implementability

Just as we gave a purely logical condition characterising the circumstances under which an objective Υ can be weakly implemented, so we would like to give a logical characterisation of the conditions under which an objective can be strongly implemented. In this section, we give a characterisation result for an important sub-class of the problem: those where the taxation function is $\{0, 1\}$. The result requires some more notation and additional intermediate results, and is considerably more involved than the equivalent result for weak implementation. Before presenting the result, we therefore give some intuition behind its structure.

Suppose we are given a game G and an objective Υ, and asked whether we can strongly implement Υ in G. In its natural state (i.e., before we impose any taxation scheme), the game G will potentially have some "good" equilibria (i.e., equilibria that satisfy Υ) and some "bad" equilibria (satisfying $\neg\Upsilon$). What we want to do is find some taxation scheme such that after imposing this scheme, all the bad equilibria are eliminated, while at least one good equilibrium remains. The way we think about this is as follows. By imposing a $\{0, 1\}$-taxation scheme on G, we will ensure that some outcome $\vec{v}$ will have no associated tax, while any choice other than this will start to incur taxes. The outcome $\vec{v}$ will serve to "tempt" players away from any potentially bad equilibria in the game by ensuring that some player has a beneficial deviation toward $\vec{v}$ in any potentially bad equilibrium, thus ensuring that such potential bad equilibria are effectively destablised. However, we also require that $\vec{v}$ does not tempt players away from all "good" equilibria (i.e., satisfying Υ); we need at least one good equilibrium to remain after the taxation scheme is imposed.

First, we have the following readily established result.

Proposition 15 *Let $G = \langle N, \Phi, c, \gamma_1, \ldots, \gamma_n, \Phi_1, \ldots, \Phi_n \rangle$, be a Boolean game, let H_G be the induced cost-free game, and let $\Upsilon \in \mathcal{L}$ be an objective. Then Υ can be strongly implemented in G iff it can be strongly implemented in H_G.*

This result indicates that restricting our attention to cost-free games is not a significant limitation.

Next, where we have a pair of valuations $\{v_i^1, v_i^2\} \subseteq \mathcal{V}_i$, we denote the *distance* between v_i^1 and v_i^2 by $\delta(v_i^1, v_i^2)$, and define this value as follows:

$$\delta(v_i^1, v_i^2) = |\{v_i^1(x) \neq v_i^2(x) \mid x \in \Phi_i\}|.$$

Thus, the distance between two valuations is the number of Boolean variables on whose value the two valuations disagree.

Where $\{v_i^1, v_i^2, v_i^3\} \subseteq \mathcal{V}_i$, we will write $(v_i^1 \sqsupseteq v_i^2 \sqsupseteq v_i^3)$ to mean that $\delta(v_i^1, v_i^3) \geq \delta(v_i^2, v_i^3)$ and $(v_i^1 \sqsupset v_i^2 \sqsupseteq v_i^3)$ to mean that $\delta(v_i^1, v_i^3) > \delta(v_i^2, v_i^3)$. The key property of this definition is given in the following:

Observation 3 *Let G be a Boolean game, let $\tau \in \mathcal{T}$ be a $\{0,1\}$-taxation scheme for G, and let $v_i^* \in \mathcal{V}_i$ be a valuation for some player i in G such that*

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

Then

$$\tau_i(v_i) = |\{p \in \Phi_i \mid v_i(p) \neq v_i^*(p)\}|$$

and for all $v_2, v_3 \in \mathcal{V}_i$, we have:

$$(v_i^2 \sqsupset v_i^3 \sqsupseteq v_i^*) \ \textit{iff} \ \tau_i(v_i^2) > \tau_i(v_i^3) \geq \tau_i(v_i^*).$$

Observation 3 establishes a link between valuations and taxation functions. Specifically, to compute the amount of tax that i incurs in a choice $v_i \in \mathcal{V}_i$, it suffices to count the number of variables on whose value v_i disagrees with the least-cost choice v_i^*. This link that allows us to establish our characterisation result:

Proposition 16 *Let G be a Boolean game with zero cost function and let $\Upsilon \in \mathcal{L}$ be an objective. Then Υ can be strongly implemented in G with a $\{0,1\}$-taxation scheme iff the following condition on G is satisfied:*

$$\exists \vec{v} : \mathit{NoBad}(\vec{v}) \ \ \& \ \ \mathit{SomeGood}(\vec{v}).$$

where:

$$\begin{array}{l}\mathit{NoBad}(v_1, \ldots, v_i, \ldots, v_n) \equiv \\ \quad \forall (v_1^1, \ldots, v_i^1, \ldots, v_n^1) \in NE_{\neg\Upsilon}(H_G) : \\ \quad\quad \exists i \in N : \exists v_i^2 \in \mathcal{V}_i : (v_i^1 \sqsupset v_i^2 \sqsupseteq v_i) \ \& \\ \quad\quad [(v_1^1, \ldots, v_i^1, \ldots, v_n^1) \models \gamma_i \leftrightarrow (v_1^1, \ldots, v_i^2, \ldots, v_n^1) \models \gamma_i)]\end{array}$$

and

$$\begin{array}{l}\mathit{SomeGood}(v_1, \ldots, v_i, \ldots, v_n) \equiv \\ \quad \exists (v_1^3, \ldots, v_i^3, \ldots, v_n^3) \in NE_{\Upsilon}(H_G) : \\ \quad\quad \forall i \in N : \forall v_i^4 \in \mathcal{V}_i : (v_i^3 \sqsupset v_i^4 \sqsupseteq v_i) \rightarrow \\ \quad\quad \neg[(v_1^3, \ldots, v_i^3, \ldots, v_n^3) \models \gamma_i \leftrightarrow (v_1^3, \ldots, v_i^4, \ldots, v_n^3) \models \gamma_i)].\end{array}$$

Proof: First note that since the game G in the problem instance has zero costs, the only beneficial deviations for a player can be to either get their goal achieved where it would not otherwise be, or to reduce taxes.

($\leftarrow$) Assume the condition on G holds: we show that this implies Υ can be strongly implemented in G. Let $\vec{v} = (v_1, \ldots, v_i, \ldots, v_n)$ be the witness to the outermost existential quantifier. Now, we define a taxation scheme τ that "fixes" the property $\varphi_{(v_1,\ldots,v_n)}$, by making all choices other than $(v_1, \ldots, v_n)$ more expensive (cf. the construction in Proposition 8). Formally:

$$\tau(p,b) = \begin{cases} 0 & \text{if } b = \vec{v}(p) \\ 1 & \text{otherwise.} \end{cases}$$

We will first show that no Nash equilibrium of $NE(G,\tau)$ satisfies $\neg\Upsilon$. For suppose that such a Nash equilibrium existed; call it $(v_1^5,\ldots,v_n^5)$. Then from Proposition 5, $(v_1^5,\ldots,v_n^5) \in NE(H_G)$, and then from the *NoBad* predicate, we have $\exists i \in N, \exists v_i^2 \in \mathcal{V}_i : (v_i^5 \sqsupset v_i^2 \sqsupseteq v_i)$ such that

$$[(v_1^5,\ldots,v_i^5,\ldots,v_n^5) \models \gamma_i \leftrightarrow (v_1^5,\ldots,v_i^2,\ldots,v_n^5) \models \gamma_i)].$$

But then player i can benefit from choosing v_i^2 instead of v_i^5, since in doing so, this would not change the status of his goal, but he would strictly reduce taxes (Observation 3). But then since some player has a beneficial deviation, then $(v_1^5,\ldots,v_n^5) \notin NE(G,\tau)$. Hence no equilibrium of (G,τ) satisfies $\neg\Upsilon$.

It remains to show that some equilibrium of (G,τ) satisfies Υ. For this, consider the *SomeGood* predicate; take an outcome $(v_1^3,\ldots,v_n^3)$ satisfying the conditions of the predicate. We claim that $(v_1^3,\ldots,v_n^3) \in NE(G,\tau)$. For suppose not; then some player has a beneficial deviation from $(v_1^3,\ldots,v_n^3)$. Since $(v_1^3,\ldots,v_n^3) \in NE_\Upsilon(H_G)$, no player can beneficially deviate to get their goal satisfied: the only beneficial deviation could be to reduce taxes. A beneficial deviation v_i^4 reducing taxes for player i must satisfy $(v_i^3 \sqsupset v_i^4 \sqsupseteq v_i)$. But then from the *SomeGood* condition, $(v_1^3,\ldots,v_i^4,\ldots,v_n^3) \not\models \gamma_i$ while $(v_1^3,\ldots,v_i^3,\ldots,v_n^3) \models \gamma_i$; but in this case, v_i^4 cannot be a beneficial deviation for i, since its goal would be unsatisfied if it chose v_i^4. We conclude that no player has a beneficial deviation from $(v_1^3,\ldots,v_n^3)$ i.e., that $(v_1^3,\ldots,v_n^3) \in NE_\Upsilon(G,\tau)$.

It follows that τ strongly implements Υ in G.

($\rightarrow$) Assume that Υ can be strongly implemented in G with a $\{0,1\}$-taxation scheme, but that the condition on games is false. Let τ be a $\{0,1\}$-taxation scheme that strongly implements Υ. Since the condition on games is false, then $\forall(v_1,\ldots,v_i,\ldots,v_n) \in \mathcal{V}_1\times\cdots\times\mathcal{V}_i\times\cdots\times\mathcal{V}_n$ one of the following conditions holds:

$$\begin{aligned}&\text{(C1)}\quad \exists(v_1^1,\ldots,v_i^1,\ldots,v_n^1) \in NE_{\neg\Upsilon}(H_G) :\\ &\qquad \forall i \in N : \forall v_i^2 \in \mathcal{V}_i : (v_i^1 \sqsupset v_i^2 \sqsupseteq v_i) \rightarrow\\ &\qquad \neg[(v_1^1,\ldots,v_i^1,\ldots,v_n^1) \models \gamma_i \leftrightarrow (v_1^1,\ldots,v_i^2,\ldots,v_n^1) \models \gamma_i)]\end{aligned}$$

or

$$\begin{aligned}&\text{(C2)}\quad \forall(v_1^3,\ldots,v_i^3,\ldots,v_n^3) \in NE_\Upsilon(H_G) :\\ &\qquad \exists i \in N : \exists v_i^4 \in \mathcal{V}_i : (v_i^3 \sqsupset v_i^4 \sqsupseteq v_i) \wedge\\ &\qquad [(v_1^3,\ldots,v_i^3,\ldots,v_n^3) \models \gamma_i \leftrightarrow (v_1^3,\ldots,v_i^4,\ldots,v_n^3) \models \gamma_i)].\end{aligned}$$

Obviously, C1 is the negation of *NoBad*, while C2 is the negation of *SomeGood*.

Now, consider the case where the universally quantified outcome $(v_1,\ldots,v_i,\ldots,v_n)$ is such that for all $i \in N$ we have

$$v_i = \arg\min_{v_i' \in \mathcal{V}_i} \tau_i(v_i').$$

Since τ is $\{0,1\}$, this outcome is unique. We reason by cases:

(C1): Suppose condition (C1) holds. Let $(v_1^1, \ldots, v_i^1, \ldots, v_n^1)$ be an outcome satisfying the conditions of (C1); note that $(v_1^1, \ldots, v_i^1, \ldots, v_n^1) \models \neg\Upsilon$. We claim that $(v_1^1, \ldots, v_i^1, \ldots, v_n^1) \in NE(G, \tau)$. For if it is not a Nash equilibrium of (G, τ), then some player i has a beneficial deviation. Since $(v_1^1, \ldots, v_i^1, \ldots, v_n^1) \in NE(H_G)$, then no player can deviate to get their goal fulfilled, so the only possible deviation would be one that would reduce taxes. We know that for all players $i \in N$ and for all choices v_i^2 such that $(v_i^1 \sqsupset v_i^2 \sqsupseteq v_i)$, we would have $(v_1^1, \ldots, v_i^1, \ldots, v_n^1) \models \gamma_i$ and $(v_1^1, \ldots, v_i^2, \ldots, v_n^1) \not\models \gamma_i$, because i cannot deviate to have her goal fulfilled.

Now, there are two possibilities:

- No outcome v_i^2 satisfies $(v_i^1 \sqsupset v_i^2 \sqsupseteq v_i)$. Then $v_i^1 = v_i$ (otherwise v_i^2 would satisfy $v_i^1 \sqsupset v_i^2 \sqsupseteq v_i)$) and in this case, v_i^1 must be a least cost choice to i, and so player i can have no beneficial deviation.
- An outcome v_i^2 exists satisfying $(v_i^1 \sqsupset v_i^2 \sqsupseteq v_i)$. In this case, we would have $(v_1^1, \ldots, v_i^1, \ldots, v_n^1) \models \gamma_i$ and $(v_1^1, \ldots, v_i^2, \ldots, v_n^1) \not\models \gamma_i$; but then v_i^2 cannot be a beneficial deviation for i.

In either case, no player can have a beneficial deviation from $(v_1^1, \ldots, v_i^1, \ldots, v_n^1)$, and so if C1 is satisfied then $(v_1^1, \ldots, v_i^1, \ldots, v_n^1) \in NE(G, \tau)$, and so, because $v^1 \models \neg\Upsilon$, τ does not strongly implement Υ in G.

(C2): Suppose condition (C2) holds. Assume $(v_1^5, \ldots, v_n^5) \in NE_\Upsilon(G, \tau)$. We will show that in fact some player has a beneficial deviation from $(v_1^5, \ldots, v_n^5)$, and hence that $(v_1^5, \ldots, v_i^5, \ldots, v_n^5) \notin NE_\Upsilon(G, \tau)$. To see this, observe that from Proposition 5, if $(v_1^5, \ldots, v_n^5) \in NE_\Upsilon(G, \tau)$ then $(v_1^5, \ldots, v_i^5, \ldots, v_n^5) \in NE_\Upsilon(H_G)$, and hence from (C2), $\exists i \in N : \exists v_i^4 \in \mathcal{V}_i$ such that $(v_i^5 \sqsupset v_i^4 \sqsupseteq v_i)$ and $[(v_1^3, \ldots, v_i^3, \ldots, v_n^3) \models \gamma_i \leftrightarrow (v_1^3, \ldots, v_i^4, \ldots, v_n^3) \models \gamma_i)]$. But then it immediately follows that player i would have a beneficial deviation (v_i^4) from $(v_1^5, \ldots, v_n^5)$, so $(v_1^5, \ldots, v_n^5) \notin NE(G, \tau)$. It follows that τ does not strongly implement Υ in G.

We therefore conclude that if Υ can be implemented in G, then the condition on games must be true.

∎

We emphasise that this characterisation is *purely logical*: the condition on games, given by the predicates $NoBad(\cdots)$ and $SomeGood(\cdots)$ do not refer to cost functions or taxation functions.

4 Secondary Criteria for Taxation Schemes

In attempting to design a taxation scheme τ for a Boolean game G, the primary aim of a principal is to design the scheme so that agents are rationally motivated to choose an outcome satisfying the objective Υ[4]. However, if it is possible to incentivise agents to satisfy Υ, then there will, in

[4]In this section, we will assume we are considering *weak* implementation, although the definitions are straightforwardly reformulated in terms of strong implementation.

general, be multiple possible taxation schemes that incentivise the agents in this way, and not all of these taxation schemes will be equally desirable. For example, imposing very high taxes might be undesirable if the same objective can be achieved with smaller taxes. In this section, therefore, we consider different *secondary* criteria that might be considered by a principal when choosing a taxation scheme; these criteria are secondary because we assume the primary aim of the principal is to ensure the satisfaction of the objective Υ. Our discussion here is somewhat inspired by the literature on welfare economics (see, e.g., [21]). We focus on the notions of *social welfare* and *equity*.

The general approach we will take is as follows. We assume the principal has some *social welfare measure*, represented by a function:

$$f : \mathcal{G} \times \mathcal{T} \times \mathcal{V}_1 \times \cdots \times \mathcal{V}_n \to \mathbb{Q}.$$

Thus, given a Boolean game G, taxation scheme τ, and outcome $\vec{v}$, a social welfare measure f gives a value $f(G, \tau, \vec{v})$, indicating how good the outcome is according to some social metric. We will assume that the aim is to maximise this function, so larger values of f are preferable. Now, in general, a scenario (G, τ) will have multiple possible outcomes (i.e., multiple possible Nash equilibria). How do we apply the function f (which evaluates individual outcomes) to such situations? The approach we adopt is by now relatively standard: we consider the value of the *worst* Nash equilibrium [16]. Formally, where f is a social welfare measure as above, then we define a two place function $\hat{f}(G, \tau)$, which gives the value of the worst Nash equilibrium of scenario (G, τ) according to f:

$$\hat{f}(G, \tau) = \begin{cases} \min\{f(G, \tau, \vec{v}) \mid \vec{v} \in NE(G, \tau)\} & \text{if } NE(G, \tau) \neq \emptyset \\ -\infty & \text{otherwise.} \end{cases}$$

Given a game G, an objective Υ, and a social welfare measure f, an *optimal* taxation scheme will be a taxation scheme $\tau^*(G, \Upsilon, f)$ that has the following properties:

1. $\tau^*(G, \Upsilon, f)$ weakly implements Υ in G;
2. $\tau^*(G, \Upsilon, f)$ maximises the value of $\hat{f}$ over all the taxation schemes satisfying (1).

If $WI(G, \Upsilon) = \emptyset$, then of course the optimal taxation scheme $\tau^*(G, \Upsilon, f)$ is undefined.

Formally, given a game G, objective Υ, and social welfare measure f, an optimal taxation scheme $\tau^*(G, \Upsilon, f)$ is defined if $WI(G, \Upsilon) \neq \emptyset$, in which case it satisfies:

$$\tau^*(G, \Upsilon, f) \in \arg \max_{\tau \in WI(G, \Upsilon)} \hat{f}(G, \tau).$$

In the subsections that follow, we will discuss several possible social welfare measures for our setting.

4.1 Social Welfare

An obvious first social welfare measure would seem to be the well-known concept of *utilitarian social welfare*. Utilitarian social welfare is conventionally defined as the sum of the utilities of players in an outcome. However, there is a difficulty when we try to consider utilitarian social welfare in our setting; namely, that utility functions $u_i(\cdots)$ are defined with respect not just to games G but also with respect to taxation functions, τ. Thus, utilitarian social welfare is, in our settings, somewhat less straightforward then for many other settings. Instead, we will here define a *qualitative* version of social welfare: the social welfare measure *qsw* simply counts the number of agents that have their goal fulfilled in a particular outcome:

$$qsw(G,\tau,\vec{v}) = |\{i \in N \mid \vec{v} \models \gamma_i\}|.$$

Thus, an optimal taxation scheme $\tau^*(G,\Upsilon,qsw)$ according to this measure will be one that maximises the number of players that have their goal satisfied in the worst case Nash equilibrium.

Another obvious alternative to measuring social welfare is to consider developing a taxation scheme that implements objective 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. Above, we defined the problem of constructing an optimal taxation scheme as a maximisation problem; we therefore express the problem of minimising tax as a maximisation problem. We thus define the social welfare measure *tb* as follows:

$$tb(G,\tau,(v_1,\ldots,v_n)) = -\sum_{i\in N} \tau_i(v_i).$$

Thus a taxation scheme $\tau^*(G,\Upsilon,tb)$, where it exists, will be one that weakly implements Υ in G, while minimising the amount of tax paid in the worst case Nash equilibrium.

It is easy to construct examples showing that minimising the total tax burden may result in outcomes that are undesirable when measured against the social welfare measure *qsw*, which we considered earlier. But nevertheless, it seems such "least intervention" approaches seem to be relatively popular in human societies.

4.2 Equity

It is, of course, well-known that an outcome which maximises (for example) utilitarian social welfare may in fact be extremely undesirable from the point of view of the majority of agents in a system. For example, the social welfare maximising outcome might allocate all the utility in the system to one agent, leaving all others with none. This motivates us to consider a range of possible other notions of equity with respect to taxation schemes, inspired to some extent by the economics literature on taxation [7].

One very obvious (although arguably naive) notion of taxation equity is to simply try to ensure that agents are taxed at broadly the same level, i.e., to minimise the maximum difference in taxes levied on different agents. Formally, the social welfare measure *md* is defined as follows:

$$md(v_1,\ldots,v_n) = -\max\{\tau_i(v_i) - \tau_j(v_j) \mid \{i,j\} \subseteq N\}$$

Simply aiming to apply the same level of taxes across an entire society may appear to be equitable, but on closer examination, it has some definite drawbacks. In particular, it does not distinguish between agents that have their goals achieved and those that do not. In the literature on taxation, the term *horizontal equity* is used to describe the idea that those in the same circumstances should be taxed at the same level [7]. 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 should only compare the taxes of agents that get their goal achieved against other agents that get their goal achieved, and we should only compare agents that do not get their goal achieved against agents that likewise do not get their goal achieved. The social welfare measure *he* will attempt to ensure that the difference in taxes paid by two agents in the same class is minimised:

$$\begin{aligned} &he(v_1,\ldots,v_n) \;= \\ &\quad -\max(\{\tau_i(v_i)-\tau_j(v_j) \mid \{i,j\}\subseteq N \;\&\; (v_1,\ldots,v_n)\models \gamma_i\wedge\gamma_j\}\cup \\ &\qquad \{\tau_i(v_i)-\tau_j(v_j) \mid \{i,j\}\subseteq N \;\&\; (v_1,\ldots,v_n)\models \neg(\gamma_i\vee\gamma_j)\}) \end{aligned}$$

5 Related Work

We now discuss how our work stands in relation to other work, both in game theory and in the multi-agent systems/algorithmic game theory community.

First, our work is very closely related to the research area known as *implementation* in game theory. The basic idea of implementation theory is that we start with a set of outcomes that we want to obtain, and we design a game so that the outcomes will result if the players in the game act rationally. We will first recall the general setting of implementation theory, and then discuss the relationship to the present paper. To keep the presentation straightforward, we will simplify the technical definitions somewhat: see [24, pp.178–180] for a more rigorous formal introduction.

We assume there is a set of game outcomes Ω, a set of players N, (as in our setting), and for each player $i \in N$ a set Ac_i of strategies/choices. Let $\mathcal{A} = Ac_1 \times \cdots \times Ac_n$ be the set of *action profiles*; we denote members of $\mathcal{A}$ by $\vec{a}, \vec{a}'$, etc. A *game form* is then a function

$$g : \mathcal{A} \to \Omega.$$

Let Γ be a set of game forms. A *preference relation* for player i is an ordering of outcomes Ω; a collection of preference relations, one for each player, is a *preference profile*. We denote preference profiles by ϖ, ϖ_1, etc. Let $\mathcal{P}$ be the set of preference profiles over N, Ω. The components $\langle N, \Omega, \mathcal{P}, \Gamma\rangle$ are together called an *environment*.

A *solution concept* is modelled as a function

$$\mathcal{S} : \Gamma \times \mathcal{P} \to \mathbf{2}^{\mathcal{A}}$$

which for every game form $g \in \Gamma$ and preference profile $\varpi \in \mathcal{P}$ define the set of action profiles $\mathcal{S}(g, \varpi)$ that may be the result of the game according to the solution concept $\mathcal{S}$.

Now, in implementation theory, a principal is assumed to have an objective defined by a *choice rule*, f, which associates a set of desirable outcomes $f(\varpi)$ with every preference profile ϖ:

$$f : \mathcal{P} \to \mathbf{2}^{\Omega}.$$

We then say a game form $g \in \Gamma$ $\mathcal{S}$-implements the choice function f if for every preference profile $\varpi \in \mathcal{P}$ we have

$$\{g(\vec{a}) \mid \vec{a} \in \mathcal{S}(g, \varpi)\} = f(\varpi).$$

In other words, the game form g $\mathcal{S}$-implements the choice function f if for every preference profile $\varpi \in \mathcal{P}$, the outcomes that could result according to solution concept $\mathcal{S}$ for the game form g and preference profile ϖ coincide with the ones that the principal would choose for this preference profile, according to the choice rule f. We could relax the equality constraint in the last equation, of course, by simply requiring that:

$$\{g(\vec{a}) \mid \vec{a} \in \mathcal{S}(g, \varpi)\} \subseteq f(\varpi).$$

Expressed as a decision problem, implementation can then be understood as follows:

> Given:
>
> - an environment $\langle N, \Omega, \mathcal{P}, \Gamma \rangle$,
> - a solution concept $\mathcal{S} : \Gamma \times \mathcal{P} \to \mathbf{2}^{\Omega}$, and
> - a choice rule $f : \mathcal{P} \to \mathbf{2}^{\Omega}$
>
> does there exist a game form $g \in \Gamma$ such that for all preference profiles $\varpi \in \mathcal{P}$, we have $\{g(\vec{a}) \mid \vec{a} \in \mathcal{S}(g, \varpi)\} = f(\varpi)$?

It should be immediately clear that the problems we are studying in the present paper are closely related to implementation problems in the sense we have just described:

- In implementation theory, preferences are given by preference relations $\succeq_i$ over outcomes. In our model, first, we make a domain restriction (preference relations are pseudo-dichotomous) and second, we use a compact representation (preferences are defined primarily by goals γ_i, expressed as propositional formulae, and secondarily by marginal costs and taxes).
- The actions Ac_i in implementation theory correspond to our choices $\mathcal{V}_i$. However, in implementation theory, the selection of a set of actions is under the control of the principal, whereas we are assuming this valuations are implicitly fixed as part of the setting.
- In implementation theory, the choice rule f can be understood as playing the role of our objective formula Υ.
- The game forms Γ that the principal may select from in implementation theory correspond to the set $\mathcal{T}$ of taxation schemes that we consider in our work. That is, in implementation theory, the principal selects a game form from Γ in order to implement a choice rule f,

while in our setting, the principal selects a taxation scheme from $\mathcal{T}$ in order to implement an objective Υ. However, in our setting the principal selects a taxation scheme while knowing the goals of players, whereas in implementation theory, a game form is selected by a principal that must induce the choice function for *all* preference profiles. In addition, in our setting, by selecting a taxation scheme, the principal is able to perturb the *preferences* of the players, while the "game form" remains unchanged.

One important sub-field of implementation theory is concerned with settings in which the preference relations of players are private (i.e., known only to the players themselves), and the actions of a player corresponds to the declaration of a preference relation by that player. Now, in such a setting, a player can in principle report any preference relation they choose: nothing obliges them to report their own actual private preferences. In this case, one can consider *incentive compatible* mechanisms: those in which a player is rationally motivated to report their true preference relation. The most important class of incentive compatible mechanisms studied in the literature are *Vickrey-Clarke-Groves* (VCG) mechanisms. VCG mechanisms have received an enormous amount of attention in the computer science literature over the past two decades [23]. Our work differs from VCG work in that we are not asking players to declare their preferences: we assume these are known.

Our work relates to a number of other topics in the multi-agent systems community and beyond. Some consideration has been given to how a principal can change the equilibrium strategies of *specific* games by introducing penalties (a form of taxation) on some actions of the players. Interesting applications include information security [28] and analyzing the TCP protocol. In the multi-agent systems community, Monderer and Tennenholtz proposed the notion of k implementation [20], whereby a principal can make payments to players (negative taxes) to incentivise players to choose certain outcomes. The setting for k-implementation is one of payments, in contrast to the present paper, and our use of Boolean games and logical objectives Υ is rather different. A related idea is discussed in [1], which considers how much compensation would have to be paid to players in a cooperative game in order for certain outcomes to become core stable.

Finally, we should mention some related work on manipulating games through communication. Grant *et al.* [12] consider a variation of Boolean games in which players have beliefs about certain variables in the game, and base their decisions about what choice to make on these beliefs. They assume a principal is able to make announcements about the value of variables, and in this way, the principal can influence the choices that players will make. The overall idea (attempting to influence the outcomes of a game) is similar to our setting, although the mechanism through which this is achieved (communication) is quite different.

6 Conclusions

In the present paper, we have studied the use of taxation schemes to incentivise behaviours in Boolean games: a natural, expressive, and compact class of logic-based games. We showed how a principal could perturb the preferences of agents in a Boolean game by imposing a taxation

scheme, and in so doing, how it could, in certain circumstances, incentivise agents to choose outcomes to satisfy an 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 fulfilled 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 questions around the question of implementing objectives Υ via taxation schemes, and also discussed some issues surrounding equitable taxation.

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, for example, techniques for the computation of taxation schemes τ for objectives Υ; and the use of taxation schemes to incentivise behaviour in richer settings, beyond the Boolean games considered in the present paper.

Acknowledgments

This paper has evolved from a paper that was published at the AAMAS-2011 conference [9]; a shorter version of the AAMAS-2011 paper was also published in the "best paper track" at the IJCAI-2011 conference. We thank the anonymous AAMAS and IJCAI referees for their useful and insightful comments. Lang thanks the project ComSoc (ANR-09-BLAN-0305-01). Wooldridge gratefully acknowledges the support of the European Research Council under Advanced Grant 291528 ("RACE").

References

[1] Y. Bachrach, E. Elkind, R. Meir, D. Pasechnik, M. Zuckerman, J. Rothe, and J. S. Rosenschein. The cost of stability in coalitional games. In *Proceedings of the 2nd International Symposium on Algorithmic Game Theory (SAGT 2009)*, pages 122–134, 2009.

[2] G. Boella, L. Sauro, and L. W. N. van der Torre. From social power to social importance. *Web Intelligence and Agent Systems*, 5(4):393–404, 2007.

[3] 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.

[4] E. Bonzon, M.C. Lagasquie, and J. Lang. Effectivity functions and efficient coalitions in boolean games. *Synthese*, 187(1):73–103, 2012.

[5] E. Bonzon, M.-C. Lagasquie-Schiex, and J. Lang. Dependencies between players in boolean games. *International Journal of Approximate Reasoning*, 50(6):899–914, 2009.

[6] E. Bonzon, M.-C. Lagasquie-Schiex, J. Lang, and B. Zanuttini. Compact preference representation and Boolean games. *Autonomous Agents and Multi-Agent Systems*, 18(1):1–35, 2009.

[7] J. J. Cordes. Horizontal equity. In R. D. Ebel J. J. Cordes and J. G. Gravelle, editors, *The Encyclopedia of Taxation and Tax Policy*. Urban Institute Press, 1999.

[8] 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.

[9] U. Endriss, S. Kraus, J. Lang, and M. Wooldridge. Designing incentives for boolean games. In *Proceedings of the Tenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2011)*, Taipei, Taiwan, 2011.

[10] 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.

[11] J. Gerbrandy. Logics of propositional control. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2006)*, pages 193–200, Hakodate, Japan, 2006.

[12] J. Grant, S. Kraus, M. Wooldridge, and I. Zuckerman. Manipulating boolean games through communication. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-2011)*, pages 210–215, 2011.

[13] D. Grossi and P. Turrini. Dependence in games and dependence games. *Autonomous Agents and Multi-Agent Systems*, 25(2):284–312, 2012.

[14] P. Harrenstein. *Logic in Conflict*. PhD thesis, Utrecht University, 2004.

[15] 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.

[16] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *Sixteenth Annual Symposium on Theoretical Aspects of Computer Science (STACS-99)*, pages 404–413, 1999.

[17] Th. Lukasiewicz and A. Ragone. A combination of boolean games with description logics for automated multi-attribute negotiation. In *Proceedings of the 22nd International Workshop on Description Logics (DL-2009)*, 2009.

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

[19] M. Mavronicolas, B. Monien, and K. W. Wagner. Weighted boolean formula games. In *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE 2007)*, pages 469–481, 2007.

[20] D. Monderer and M. Tennenholtz. K-implementation. *Journal of AI Research*, 21:37–62, 2004.

[21] H. Moulin. *Axioms of Cooperative Decision Making*. Cambridge University Press: Cambridge, England, 1988.

[22] 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.

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

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

[25] C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley: Reading, MA, 1994.

[26] L. Sauro and S. Villata. Dependency in cooperative boolean games. *Journal of Logic and Computation*, 2012.

[27] J. Sloman and A. Wride. *Economics (7th edition)*. Prentice Hall, 2009.

[28] W. Sun, X. Kong, D. He, and X. You. Information Security Game Analysis with Penalty Parameter. In *Proceedings of the International Symposium on Electronic Commerce and Security (ISECS–2008)*, pages 453–456, 2008.

[29] J. Tobin. A proposal for international monetary reform. *Eastern Economic Journal*, 4(3–4):153–159, 1978.

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