ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-6

Bad equilibria (and what to do about them)

Michael Wooldridge[1]

Abstract. I begin by arguing that the notion of *economic equilibrium* is an important analytical tool with which to understand the behaviour of today's networked computer systems. This is because the behaviours that such systems exhibit are in part a function of the preferences and desires of system participants; this gives such systems the flavour of an economic system. In economics, an equilibrium is a steady-state situation, which obtains because no participant has any rational incentive to deviate from it. Equilibrium concepts are arguably the most important and widely used analytical weapons in the game theory arsenal. The concept of Nash equilibrium in particular has found a huge range of applications, in areas as diverse and seemingly unrelated as evolutionary biology and moral philosophy. However, there remain fundamental problems associated with Nash equilibria and their application, which must be considered if we want to apply them to the analysis of computer systems. First, there may be multiple Nash equilibria, in which case, how should we choose between them? Second, some equilibria may be undesirable, in which case, how can we avoid them? In this essay, I will introduce work that we have done addressing these problems from a computational/AI perspective. Assuming no prior knowledge of game theory or economic solution concepts, I will discuss various ways in which we can try to engineer a scenario so that desirable equilibria result, or else engineer out undesirable equilibria.

1 Introduction

My primary aims in this paper are twofold:

1. First, I want to argue that the notion of *economic equilibrium* is an important concept through which to understand today's networked computer systems. In particular, I argue that economic/game theoretic equilibrium concepts are of potential value for understanding systems such as the Internet.
2. Second, I want to describe (in outline only) some work that we have done on *managing* the equilibria of systems. In particular, I will describe how we can perturb the behaviour of rational agents so that they will select equilibria that satisfy certain logically-specified properties. The mechanism we consider, through which such manipulation can be achieved, is taxation.

In economics, an equilibrium is nothing more than a steady-state situation, which obtains and persists because no participant has any rational incentive to deviate from it. Equilibrium concepts are the most important and widely used analytical weapons in the game theory arsenal [15]. The concept of Nash equilibrium in particular has found a huge range of applications, in areas as diverse and seemingly unrelated as evolutionary biology and moral philosophy. The first main argument of this paper is that the standard analytical tools developed within computer science over the past four decades will be inadequate and inappropriate for understanding the kinds of behaviours that might be exhibited by complex distributed systems such as the Internet. This is because the overall behaviour of such systems cannot be understood without reference to the fact that the components of the system are not altruistic or even benevolent: the agents on the Internet will typically act in their own interests, as far as they see them. It is commonly accepted in economics and game theory that the notion of equilibrium is appropriate for analysing systems containing multiple self-interested actors; I will argue that the notion of equilibrium is also of value to understanding systems such as the Internet. I will mainly focus on the notion of *Nash equilibrium*, which is the best-known and most important equilibrium concept in game theory.

I will then highlight some issues that arise if we want to apply the concept of Nash equilibrium to understanding distributed systems:

- First, there may be multiple Nash equilibria, in which case, how should one of them be chosen? How can we decide which will actually result?
- Second, some equilibria of the system may be inefficient or otherwise undesirable – in which case, what can we do to avoid these equilibria? What interventions are available to steer the system towards more desirable equilibria?

Following this discussion, I will introduce some work that we have done addressing these problems from a computational/AI perspective. Assuming no prior knowledge of game theory or economic solution concepts, I will discuss how we can try to engineer a scenario so that desirable equilibria result, or else engineer out undesirable equilibria. I will focus on the idea of imposing *taxation schemes* on systems, so that the preferences of rational agents are perturbed in such a way that the components of the system will choose a desirable outcome in equilibrium.

2 Setting the Scene

It is a well-established scientific tradition that any invited paper worth its salt should attempt to pass off a number of hoary cliches as if they were profound and original insights. I have no wish to offend tradition, so let me get my cliches in early:

1. *The future of computing will be one of ubiquitous, seamlessly interconnected computing devices.*
2. *These devices will be increasingly sophisticated and mobile.*
3. *We will continue to delegate ever more tasks to these devices as part of our everyday lives.*
4. *The development of techniques for modelling, programming, and analysing such systems represents one of the key chal-*

[1] Dept of Computer Science, University of Oxford, Oxford OX1 3QD, UK. Email mjw@cs.ox.ac.uk.

lenges for computer science in the early part of the 21st century.

Many trends in contemporary computing are a reflection of these self-evident truths. Examples include the semantic web [3], ubiquitous computing [12], autonomic computing [14], cloud computing, and my own research field, multi-agent systems [19]. It is by now generally accepted that classical computational models (such as the Turing machine), and the associated theory that goes with these models, is not really appropriate for modelling and understanding such systems. Much of the energy and effort of the computing research community over the past three decades has been directed to developing alternative models, programming languages, and theories, through which we can better develop and understand such systems; notable examples of such work include process algebras such as Milner's π calculus [13].

Now, I am going to claim that the notion of *economic equilibrium* is an appropriate concept through which we can understand and analyse an important class of such systems. To understand my argument, let us first recall a well-known paradigm for program development, based around the notion of *program correctness*. This paradigm has underpinned much computer science research since the 1960s. Roughly speaking, the story of program correctness is usually told as follows. We start with a *specifier*, who constructs a specification for a program. In simple terms, this specification describes what the program should do. We then build a program, and we check it against its specification. The program is judged to be correct if its satisfies (meets, fulfils) its specification; otherwise it is incorrect. Typically we write $P \models \varphi$ to mean that program P satisfies the specification φ. A little more formally, the idea is usually that the formal specification φ defines a set $[\![\varphi]\!]$ of *behaviours* – the behaviours of the program that are deemed to be acceptable. A program P is also associated with a set $[\![P]\!]$ of behaviours; these are the possible runs, or computations of the program. Then the program P is said to be correct with respect to the specification φ if $[\![P]\!] \subseteq [\![\varphi]\!]$. This is pretty much the story as told in the temporal verification of computer programs [10, 11], and the associated technology of model checking [5]: model checking, for example, is essentially concerned with the problem of checking whether $[\![P]\!] \subseteq [\![\varphi]\!]$, where φ is expressed as a temporal logic formula.

Now let us step back from this story a little. Notice that in this very well-known story, it is assumed that there is somebody – the specifier – who is in what we might call a privileged position. That is, the specifier defines the specification, and thus has complete authority to say what is "correct" behaviour and what is "incorrect" behaviour for the system under question. Only programs P that satisfy the specification φ are deemed to be acceptable. (Of course, whether the specification is drawn up by a committee or by an individual is not really relevant; the point is that there is a single standard of behaviour, defined by φ, and anything in contradiction with this judged to be an error.)

Now consider this paradigm applied to systems such as the Internet. Does it make sense? In one sense, certainly. For example, standards bodies such as the World-Wide Web Consortium (W3C) and the Internet Engineering Task Force (IETF) define the protocols underpinning the Internet, and we can of course check whether these protocols are being correctly implemented, in the sense that the correct packet types are being sent in the correct order, and that responses of the correct type are being given at the appropriate juncture. But such an analysis, important though it may be, is really missing a very big part of the story that is relevant to understanding how the Internet behaves. The point is that the classical view of correctness assumes *a single standpoint of correctness*. But with systems like the Internet, *nobody* is in such a privileged position. With my tongue firmly in cheek for a moment, consider that the W3C and IETF in all likelihood deplore the fact that teenagers use the Internet to illegally download music and videos; they are surely horrified by the fact that terrorists use the Internet to communicate and coordinate their attacks; and they may be deeply disapproving of the ocean of pornography that washes across the Internet every day; but these concerns are nothing to do with the Internet being correct or otherwise. Asking "Is the Internet correct" does not make sense. Indeed, the question is a category error, in the same way that asking "is 9 o'clock green" is a category error. The question is meaningless. To apply the concept of correctness, there must be a privileged position, from which a unique standard of correctness φ may be prescribed. And in systems like the Internet, there is and can be no such privileged position.

There is no privileged standpoint of correctness on the Internet because the millions (and soon, billions) of players in the Internet are not acting on behalf a single individual or organisation. Nor are they benevolent or selflessly altruistic entities. They use the Internet to further their own ends: they are *self interested*, and they will, if necessary, *act strategically* in order to obtain the best outcome for themselves in any given situation. Thus trying to understand a system like the Internet in terms of the packet-level protocols exchanges that take place is irrelevant if we want to understand its higher-level dynamics. The systems involved can and should be understood not just as a network of computer processors exchanging data streams according to certain protocols, but as *computational economic systems*. If we ignore self-interest, strategic, and economic considerations when we conceptualise and design such a system, then we will be ignoring and missing issues that are *fundamental* in order to understand the likely behaviours of the system.

To take a specific example, consider eBay, the online auction house. When users create an auction on eBay, they must specify a deadline for bidding in the auction. This deadline, coupled with the strategic concerns of bidders, leads to behaviour known as sniping [16]. Roughly, sniping is where bidders try to wait for the last possible moment to submit bids. Sniping is *strategic behaviour*, used by participants to try to get the best outcome for themselves. It is perhaps the best-known behaviour that is witnessed on eBay. Now, in one sense, it is perfectly coherent to try to model and analyse the protocols and system structure of eBay using existing techniques for the analysis of distributed systems; but this analysis will not predict or explain sniping, for the simple reason that such analyses do not take self-interest or strategic considerations into consideration. Thus, to understand the likely trajectories of a system such as eBay, *we have to take into account its nature as a computational economy, populated by self-interested agents acting strategically*. If we do not take into account preferences/goals and strategic behaviour, then we largely miss the point of a system like eBay; and in the case of eBay, we won't be able to predict or understand its most characteristic behaviour – sniping.

So, if we cannot apply the concept of correctness systems like the Internet, what can we use instead? I argue that we can usefully apply the concept of *equilibrium*. In its everyday sense, the term equilibrium simply means a steady state situation, in which opposing forces are balanced in such a way as to maintain the steady state. In economics, the forces in question are the preferences or desires of those participating in the scenario. An economic equilibrium is a steady state condition that obtains because no participant is rationally mo-

tivated to deviate from it. So, my argument is that instead of asking "does the system satisfy my specification", we need to ask "what are the *equilibria* of the system, and what properties do these equilibria have?" Such an analysis has, I believe, a better chance of being able to predict and understand properties such as sniping on eBay than other more conventional analytical concepts (such as analyses based around the notion of correctness), simply because it takes into account the fact that the participants are self-interested.

Game theory uses a number of models to try to capture scenarios containing multiple self-interested agents, and considers a range of equilibrium concepts [15]. My aim here is not to present a detailed study of these models, but to hint at their key components, and to indicate how they relate to computational systems such as the Internet.

In game theory, a "game" is a model of a situation in which self-interested agents interact. Typically, a game specifies:

- The participants in the system (the "players" of the game).
- The beliefs that the participants have, about the other players of the game and the state of the world.
- The possible choices/actions/strategies available to each of the agents in the system.
- The effect that each combination of choices has.
- The preferences that each agent in the system has over each possible outcome.

A key concern in game theory is to try to understand what the outcomes of a game can or should be, under the assumption that the players within it act rationally. To this end, a number of *solution concepts* have been proposed, of which Nash equilibrium is perhaps the best-known. A Nash equilibrium is a collection of choices such that no player can benefit by unilaterally deviating from this combination of choices. Nash equilibria seem like reasonable candidates for the outcome of a game because to move away from a Nash equilibrium would result in some player being worse off, which clearly seems to be irrational.

At a high level, it seems fairly straightforward to understand computer systems in terms of these concepts: the players map to non-deterministic programs (i.e., programs that have choices, e.g., about what message to send next), and actions/strategies map to the choices available to programs. The preferences can be assumed to be the preferences of the individual on whose behalf the program is acting. In eBay terms, the agents will be the seller and the various bidders. The seller prefers a high price; buyers prefer a low price.

Once we can formulate a system as a game in this way, we can start to ask, for example, what its equilibria are, and whether they are desirable. I am of course glossing over a whole raft of issues that need to be addressed to make such an analysis work; and I expect these issues to drive much research over the next few years. But here I want to draw attention to just one issue: In general, a system can have *undesirable* equilibria. For example, the system can have "inefficient" equilibria, where the rational outcome that results could be improved for *everybody* (this is the case in the famous *Prisoner's Dilemma* [2]). In this case, what interventions are available that can help to steer the system towards more desirable equilibria? In our work, we have explored several possibilities. For example, one can use *communication* to alter the beliefs of system participants [8]. By altering their beliefs (the basis on which they choose their actions), we can perturb the system towards more desirable outcomes than would otherwise be chosen. Another possibility is to declare "laws" – that is, to define sets of rules or standards or behaviour that agents are expected to adhere to. In multi-agent systems research, this is the domain of *social laws* [17]. A common problem with social laws (in human and artificial societies) is that of *compliance*: why should a rational agent comply with a set of rules when it is not in their interests? One possibility is to try to construct social laws such that compliance is in the interest of all concerned [1]. In general, of course, this will not always be possible. Much of the remainder of this paper is taken up with another possibility: the idea of overlaying systems with *taxation schemes*, so that the actions of agents are taxed in various ways depending on the choices they make. If we design the taxation scheme appropriately, we can perturb the preferences of the participants away from undesirable equilibria, towards more desirable ones. The next section presents this work in more detail, and also serves as an exemplar of the general kind of framework in which we can study these problems [7].

3 Incentivising Desirable Equilibria

In this section, I will move away from the abstract discussion presented above, and give a concrete example of work that we have done that was driven by the considerations presented above. The work addresses the problem of how to deal with "bad equilibria" – equilibria that are judged to be undesirable for some reason, often because they are inefficient. The work uses the model of *Boolean games*.

Boolean games are a natural, expressive, and compact class of games, based on propositional logic; and they have a natural computational interpretation. Boolean games were introduced in [9], and their computational and logical properties have subsequently been studied by several researchers [4, 6]. In such a game, each agent i is assumed to have a goal, represented as a propositional formula γ_i over some set of variables Φ. In addition, each agent i is allocated some subset Φ_i of the variables Φ, with the idea being that the variables Φ_i are under the unique control of agent i. The choices, or strategies, available to i correspond to all the possible allocations of truth or falsity to the variables Φ_i. An agent will try to choose an allocation so as to satisfy its goal γ_i. Strategic concerns arise because whether i's goal is in fact satisfied will depend on the choices made by others.

We introduce the idea of imposing taxation schemes on Boolean games, so that a player's possible choices are taxed in different ways. Taxation schemes are designed by an agent external to the game known as the *principal*. The ability to impose taxation schemes enables the principal to *perturb the preferences of the players in certain ways*: all other things being equal, an agent will prefer to make a choice that minimises taxes. As discussed above, the principal is assumed to be introducing a taxation scheme so as to incentivise agents to achieve a certain desirable outcome; or to incentivise agents to rule out certain undesirable outcomes. We represent the outcome that the principal desires to achieve via a propositional formula Υ: thus, the idea is that the principal will impose a taxation scheme so that agents are rationally incentivised to make individual choices so as to collectively satisfy Υ. However, a fundamentally important assumption in what follows is that taxes do not give us absolute control over an agent's preferences. In our setting specifically, it is assumed that no matter what the level of taxes, *an agent would still prefer to have its goal achieved than not*. This imposes a fundamental limit on the extent to which an agent's preferences can be perturbed by taxation.

We begin in the following section by introducing the model of Boolean games that we use throughout the remainder of the paper. We then introduce taxation schemes and the *incentive design problem* – the problem of designing *taxation schemes* that will influence the behaviour of agents within a game so that they will act so as to satisfy

a certain logically-specified objective Υ in equilibrium.

Propositional Logic: Let $\mathbb{B} = \{\top, \bot\}$ be the set of Boolean truth values, with "$\top$" being truth and "$\bot$" being falsity. 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. Let $\mathcal{V}$ denote the set of all valuations over Φ. We write $\models \varphi$ to mean that φ is a tautology. We denote the fact that $\varphi, \psi \in \mathcal{L}$ are logically equivalent by $\varphi \Leftrightarrow \psi$; thus $\varphi \Leftrightarrow \psi$ means that $\models \varphi \leftrightarrow \psi$.

Agents, Goals, and Controlled Variables: The games we consider are populated by a set $Ag = \{1, \ldots, n\}$ of *agents* – the players of the game. Think of these as the components of a distributed system. Each agent is assumed to have a *goal*, characterised by an $\mathcal{L}$-formula: we write γ_i to denote the goal of agent $i \in Ag$. Each agent $i \in Ag$ *controls* a (possibly empty) subset Φ_i of the overall set of Boolean variables (cf. [18]). By "control", we mean that i has the unique ability within the game to set the value (either $\top$ or $\bot$) of each variable $p \in \Phi_i$. We will require that $\Phi_1, \ldots, \Phi_n$ forms a partition of Φ, i.e., every variable is controlled by some agent and no variable is controlled by more than one agent ($\Phi_i \cap \Phi_j = \emptyset$ for $i \neq j$). Where $i \in Ag$, a *choice* for agent i is defined by a function $v_i : \Phi_i \rightarrow \mathbb{B}$, i.e., an allocation of truth or falsity to all the variables under i's control. Let $\mathcal{V}_i$ denote the set of choices for agent i. The intuitive interpretation we give to $\mathcal{V}_i$ is that it defines the *actions* or *strategies* available to agent i; the *choices* available to the agent. Thus, we can think of an agent i as a non-deterministic program, which can assign values to its variables Φ_i as it chooses.

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

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

This notion of a cost function represents an obvious generalisation of previous presentations of Boolean games: costs were not considered in the original presentation of Boolean games [9, 4], and while costs were introduced in [6], it was assumed that only the action of setting a variable to $\top$ would incur a cost. In fact, as we discuss in the parent paper, costs are, in a technical sense, not required in our framework; we can capture the key strategic issues at stake without them. This is because we can "simulate" marginal costs with taxes. However, it is natural from the point of view of modelling to have costs for actions, and to think about costs as being imposed from within the game, and taxes, (defined below), as being imposed from without.

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

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

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

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

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

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

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

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

Taxation Schemes: A taxation scheme defines additional (imposed)

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

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

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

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

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

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

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

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

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

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

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

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

This definition has the following properties:

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

Solution Concepts: Given this formal definition of utility, we can define solution concepts in the standard game-theoretic way [15]. In this paper, we focus on (pure) Nash equilibrium. (Of course, other solution concepts, such as dominant strategy equilibria, might also be considered, but for simplicity, in this paper we focus on Nash equilibria.) We say an outcome $(v_1, \ldots, v_i, \ldots, v_n)$ is a Nash equilibrium if for all agents $i \in Ag$, there is no $v_i' \in \mathcal{V}_i$ such that $u_i(v_1, \ldots, v_i', \ldots, v_n) > u_i(v_1, \ldots, v_i, \ldots, v_n)$. Let $NE(G, \tau)$ denote the set of all Nash equilibria of the game G with taxation scheme τ.

Incentive Design: We now come to the main problems that we consider in the remainder of the paper. Suppose we have an agent, which we will call the principal, who is external to a game G. The principal is at liberty to impose taxation schemes on the game G. It will not do this for no reason, however: it does it because it wants to provide incentives for the agents in G to choose certain collective outcomes. Specifically, the principal wants to incentivise the players in G to choose rationally a collective outcome that satisfies an *objective*, which is represented as a propositional formula Υ over the variables Φ of G. We refer to this general problem – trying to find a taxation scheme that will incentivise players to choose rationally a collective outcome that satisfies a propositional formula Υ – as the *implementation problem*.

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

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

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

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

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

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

So, what objectives Υ can be weakly implemented? At first sight, it might appear that the satisfiability of Υ is a sufficient condition

for implementability. Consider the following naive approach for constructing taxation schemes with the aim of implementing satisfiable objectives Υ:

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

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

Proposition 1 *There exist instances of the* WEAK IMPLEMENTATION *problem with satisfiable objectives Υ that cannot be weakly implemented.*

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

Proposition 2 *There exist instances of the* WEAK IMPLEMENTATION *problem with tautologous objectives Υ that cannot be implemented.*

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

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

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

Proposition 3 *For all games G and objectives Υ, if the formula Υ' is satisfiable:*

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

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

4 Conclusions

I believe that the notion of economic equilibrium has an important role to play in the analysis of today's networked computer systems. In this paper I have tried to explain why I believe this, and to sketch out some of the issues that arise if we take this idea seriously. The grand challenge underpinning this work is to develop techniques that will enable us to analyse, understand, and predict the behaviour of computer systems when the participants in these systems are self-interested; and to be able to manage the equilibria of such systems. The issues raised by this work seem to be highly relevant for computer science, conceptually interesting, and technically deep: surely an irresistible combination. **Acknowledgments:** This research was supported by the European Research Council under Advanced Grant 291528 ("RACE"). I have benefited enormously from discussions with Rahul Savani. Part of the research reported in this paper was carried out with jointly with Ulle Endriss, Sarit Kraus, and Jérôme Lang.

REFERENCES

[1] T. Ågotnes, W. van der Hoek, and M. Wooldridge. Normative system games. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2007)*, Honolulu, Hawaii, 2007.

[2] R. Axelrod. *The Evolution of Cooperation*. Basic Books: New York, 1984.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, May 2001.

[4] 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.

[5] E. M. Clarke, O. Grumberg, and D. A. Peled. *Model Checking*. The MIT Press: Cambridge, MA, 2000.

[6] 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.

[7] 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.

[8] J. Grant, S. Kraus, M. Wooldridge, and I. Zuckerman. Manipulating boolean games through communication. In *Proceedings of the Twenty Second International Joint Conference on Artificial Intelligence (IJCAI-2011)*, Barcelona, Catalonia, Spain, 2011.

[9] 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.

[10] Z. Manna and A. Pnueli. *The Temporal Logic of Reactive and Concurrent Systems*. Springer-Verlag: Berlin, Germany, 1992.

[11] Z. Manna and A. Pnueli. *Temporal Verification of Reactive Systems — Safety*. Springer-Verlag: Berlin, Germany, 1995.

[12] R. Milner. Ubiquitous computing: Shall we understand it? *The Computer Journal*, 49(4):383–389, 2006.

[13] R. Milner, J. Parrow, and D. Walker. A calculus of mobile processes. *Information and Computation*, 100(1):1–77, September 1992.

[14] R. Murch. *Autonomic Computing*. IBM Press, 2004.

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

[16] A. Roth and A. Ockenfels. Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the internet. *American Economic Review*, 92(4):10931103, 2002.

[17] Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, San Diego, CA, 1992.

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

[19] M. Wooldridge. *An Introduction to Multiagent Systems (second edition)*. John Wiley & Sons, 2009.