Econ Theory Bull (2017) 5:9–19
DOI 10.1007/s40505-016-0107-y

CrossMark

RESEARCH ARTICLE

A folk theorem with codes of conduct and communication

Juan I. Block[1] · **David K. Levine**[2,3]

Received: 17 August 2016 / Accepted: 3 September 2016 / Published online: 16 September 2016

Abstract We study self-referential games in which players have the ability to commit to a code of conduct—a complete description of how they play and their opponents should play. Each player receives a private signal about each others' code of conduct and their codes of conduct specify how to react to these signals. When only some players receive informative signals, players are allowed to communicate using public messages. Our characterization of the effect of communication on the equilibrium payoffs yields a folk theorem and players share their private information truthfully in equilibrium. We also provide an application of codes of conduct: games that are played through computer programs.

Keywords Folk theorem · Self-referential game · Commitment · Communication

JEL Classification C70 · D03

1 Introduction

In many economic environments people develop social norms or conventions when they regularly play familiar (or similar classes of) games, specifying behavior and choices everyone is expected to conform to. These social norms naturally endow

✉ Juan I. Block
jb2002@cam.ac.uk

[1] Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, UK

[2] Department of Economics, European University Institute, Villa San Paolo, Via della Piazzuola 43, 50133 Firenze, Italy

[3] Department of Economics, Washington University in St. Louis, Campus Box 1208, St. Louis, MO 63130-4899, USA

people with the ability to commit, and this in turn allows the possibility that certain behavior can be visible to other people due to costs of choosing alternatives outside the conventional norm (see, for example, Schelling 1960; Frank 1988). Such consensus on behavior gives rise to the emergence of cooperation even when these economic relationships are not sustained in the long run, and therefore, not susceptible to future incentives.

In this paper, we examine such situations where players employ codes of conduct which are defined as a complete specification of how they play and their opponents should play. Players also receive private signals about what code of conduct their opponents may be using, while their own code of conduct enables them to respond to these signals. We focus on the limit case of perfectly informative signals because we are interested in applications such as games played through agents whose codes of conduct determine their compensation schemes (the contracts players sign with their agents are observable), if the agent is human, or are embedded in their programming, if automated.

We show a folk theorem for finite normal form games using simple trigger codes of conduct and under two observability assumptions. First, we make the relatively standard assumption that all players can observe their opponents' codes of conduct as in many models of conditional commitment devices (see, for example, Tennenholtz 2004; Kalai et al. 2010). We demonstrate that our codes of conduct generalize the commitment device space described by Kalai et al. (2010) and we discuss how codes of conduct can be applied to, but are not restricted to computer algorithms (program equilibrium as described in Tennenholtz (2004)). These program strategies are self-referential in the sense that they take as input the opponents' programs; then they syntactically compare them with its own description and execute an output strategy depending whether they are equal or not.

Our main result, however, is for settings in which not all players can observe opponents' codes of conduct. In large communities, for example, a group of people can be excluded from monitoring or have limited ability to screen whether other people would behave within the current convention. In general, it may be reasonable to assume that people have good information about whether a few people with whom they closely interact use the same code of conduct that they do, but not so reasonable that they would have this information about everyone in the community. However, we allow the possibility that people can communicate via cheap talk what they have observed about others. Specifically, we extend the idea of self-referential games to assume that after observing private signals about rivals' codes of conduct players are allowed to send public "cheap talk" messages. Communication is per se costless but within the self-referential framework we show that it becomes a "signaling" device since the code of conduct includes the cheap talk messages, thereby players may receive different private signals depending on different message strategies. Although players choose strategically what to communicate, we show that there is an equilibrium in which players will reveal their private information truthfully. As a result, our key finding is that with public communication it is sufficient to get a folk theorem that every player is observed by at least two other players.

1.1 Related literature

The key feature that distinguishes our paper from conditional commitment device models is that our commitment device (code of conduct) is a function of signals rather than the opponents' devices. The two most closely related papers are Tennenholtz (2004) and Kalai et al. (2010) in which agents commit to a particular behavior in response to their opponents' device. By contrast, in our model the commitment device affects the likelihood of signals on which the other players would base their behavior; it requires an indirect connection between commitment and observation, thereby relying on a weaker observability assumption and expanding the number of applications our model can be used for. Our model encompasses Tennenholtz's model of conditional commitment devices where players play through computer programs and receive as input their opponents' program. Unlike Tennenholtz, we fully describe the space of codes of conduct and our folk theorem reaches efficiency. Similar to Kalai et al. (2010), we characterize the space of codes of conduct avoiding typical circularity problems, but in contrast to their model, we allow for mixed strategies, do not require the use of jointly controlled lotteries, and consider normal form games with more than two players. In their models, every player condition her play on all the other players' devices, and their equilibrium construction breaks down if this observability assumption is not satisfied. Unlike these papers, our main focus is on situations where players are able to observe some opponents' code of conduct, and we propose a theoretical approach for incorporating public communication via cheap talk messages into conditional commitment device models.

More recently, attention has been drawn to noisy environments; Block and Levine (2015) examine players that observe imperfectly informative signals about each others' codes of conduct, as Levine and Pesendorfer (2007) do within a evolutionary framework. Block and Levine (2015) prove a folk theorem for repeated games with private monitoring. How the period at which players receive signals about others' codes of conduct affects the equilibrium payoff set was explored by Block (2013). In a much less noisy environment, Bachi et al. (2014) study games in which deceptive players may betray their true intentions. They show a folk theorem for two-player normal form games if the cost of deception is sufficiently low. By contrast, in our model these kinds of costs are embedded in the likelihood of private signals but could be readily incorporated. Although a leading example to motivate commitment in this literature is the idea of a communication phase before the actual play of the underlying game, to the best of our knowledge this is the first paper that incorporates explicitly such a communication round.

2 The model

2.1 The baseline game

We study a finite N-player normal form game $\Gamma = (I, (S_i, u_i)_{i \in I})$. There is a set of N players $I = \{1, 2, \dots, N\}$ indexed by i. Each player i chooses a strategy s_i from the finite strategy set S_i. Let $S = \times_i S_i$ be the product space of the individual strategy sets. Let $s \in S$ denote a strategy profile. The payoff of player i is $u_i : S \to \mathbb{R}$.

We write $\Delta(S_i)$ for the set of mixed strategies for player i and let $\Delta(S) = \times_i \Delta(S_i)$. To avoid dealing with measure theoretic considerations when we describe the self-referential framework, we restrict attention to a finite subset of those mixed strategies for each player i that we denote by $\mathsf{S}_i \subseteq \Delta(S_i)$ with generic element σ_i. We extend payoffs to mixed strategy profiles $\sigma \in \mathsf{S} = \times_i \mathsf{S}_i$ in the standard way. Define a minmax strategy against player i as $\underline{\sigma}^i_{-i} \in \arg\min_{\sigma_{-i} \in \mathsf{S}_{-i}} \max_{\sigma_i \in \mathsf{S}_i} u_i(\sigma_i, \sigma_{-i})$. Let $\underline{u}_i = u_i(\underline{\sigma}^i_i, \underline{\sigma}^i_{-i})$ be the minmax value of player i where $\underline{\sigma}^i_i$ denote i's best response to $\underline{\sigma}^i_{-i}$.

2.2 The self-referential game

We present the model and notation that we introduced in Block and Levine (2015). For any baseline game Γ, we embed Γ in the self-referential framework, and therefore define the self-referential game $G(\Gamma)$. In the beginning of the game G, every player i privately observes a signal $y_i \in Y_i$, where Y_i is finite. Let $y \in Y = \times_i Y_i$ be a private signal profile. A strategy for player i in G is a mapping from his set of signals to his subset of strategies in Γ and a mapping for each other player from their set of signals to their subset of strategies in Γ. We referred to such strategy as a code of conduct denoted by r_i. We emphasize that a code of conduct for each single player specifies how all players should play. The reason why we assume that a code of conduct specifies what the player expects from others is that, first, this is how the majority of social norms are built and, second, it allows us to have a well-defined notion of agreement between players when the game is asymmetric and players have different roles.[1] Formally, a code of conduct r_i is a $1 \times N$ vector for which each coordinate j corresponds to a mapping from a set of signals Y_j to the subset of mixed strategies S_j.[2] We think of codes of conduct as social norms that emerge when people interact in familiar games. Specifically, each player chooses a code of conduct that everyone is supposed to follow (what to play conditional on each private signal y_j) and simultaneously commits to follow the code himself. We assume that there is a common space of codes of conduct

$$R_0 = \Big\{ r_i = (r_i^j)_{j \in I} : \forall j \in I, r_i^j \in \mathsf{S}_j^{Y_j} \text{ and } \\ \forall y_j \in Y_j, r_i^j(y_j) \in \mathsf{S}_j \Big\},$$

[1] Note that in symmetric two player games there is no ambiguity in referring to observing an informative signals that may reveal whether the opponent is using the same strategy or not, but if the game is asymmetric the notion of "same strategy" can be captured with codes of conduct as players may agree on what each player should do.

[2] Similar to Tennenholtz (2004), but unlike Kalai et al. (2010), codes of conduct allow players to randomize among their Γ strategies.

where $\mathsf{S}_j^{Y_j}$ represents the set of all mappings with domain Y_j and range S_j.[3] For sake of simplicity and to avoid existence issues, we assume that the minmax strategies and any mixed strategy equilibrium of the underlying game Γ belongs to R_0.[4] We write $r \in R = \times_i R_0$ for the profile of codes of conduct. With some abuse of notation, we write $r_i(y_i) = r_i^i(y_i)$.

Crucial to codes of conduct is the ability of players to receive signals about the codes of conduct used by opponents. We model this by assuming that for each $r \in R$ there is a probability distribution $\pi(\cdot|r) \in \Delta(Y)$ over Y. We let $\pi_i(\cdot|r)$ denote the marginal probability distribution of $\pi(\cdot|r)$ over Y_i. For any $y_i \in Y_i$, let $\pi_i(y_i|r)$ be the probability of y_i given $r \in R$.

We can now define the expected payoff from using codes of conduct: for player i is

$$U_i(r) = \sum_{y \in Y} u_i(r_1(y_1), \ldots, r_N(y_N))\pi(y|r).$$

Note that codes of conduct determine both how players behave as a function of the signals they receive and the probability distribution over the signals players receive about each others' codes of conduct. As a result, the expected payoff of player i can be decomposed into two parts: the first part u_i depends on the actual play, that is, $r_i(y_i) = \sigma_i$ for each $i \in I$; while the second part $\pi(y|r)$ is determined by both what players planned to choose and what they expected from their opponents, that is, $r = ((r_1^j)_{j \in I}, \ldots, (r_N^j)_{j \in I})$.

The timeline of the self-referential game is as follows:

1. Each player simultaneously chooses r_i, which is not observed by the other players.
2. Each player privately observes the realization y_i of his own signal.
3. Each player chooses a strategy σ_i according to r_i.

A Nash equilibrium in the self-referential game G (or a self-referential equilibrium) is a profile of codes of conduct $r \in R$ such that for all players i and any $\tilde{r}_i \neq r_i$, it follows

$$\sum_{y \in Y} u_i(r_i(y_i), r_{-i}(y_{-i}))\pi(y|r) \geq \sum_{y \in Y} u_i(\tilde{r}_i(y_i), r_{-i}(y_{-i}))\pi(y|\tilde{r}_i, r_{-i}).$$

[3] In contrast to Kalai et al. (2010), we assume that players do not randomize when they are choosing their codes of conduct, that is, a mixed code of conduct $\rho_i \in \Delta(R_0)$. However, this would not make a difference since the probability distribution over signals depends on the actual code of conduct profile r. In their setting it matters because players condition directly on their opponents' conditional devices so after the randomization players know their opponents' choice. To avoid this difficulty, their space of commitment devices prohibits some responses to pure choices.

[4] Notice that players could always ignore the signals and play Γ, and this in turn implies that they can attain any equilibrium of Γ. However, our codes of conduct space does not satisfy the "voluntary" condition proposed by Kalai et al. (2010) because although players can choose without conditioning on the signal they observe (what they called "unconditioned play") this might still induce different signal distributions and result in some conditioning by their opponents, therefore violating their "private play" condition.

3 The folk theorem

Social norms are essentially based on a broad notion of "reciprocity," meaning that people conform and expect others to conform, and people would conform if all others conform. We proceed to define a self-referential framework within which social norms might be such that if people are likely to act according to the current social norm, then alternative behavior should be visible to other people. We first assume that each player is able to detect all the opponents that do not choose the same code of conduct prescribed by the conventional norm. In other words, players can directly observe their opponents' codes of conduct as in Tennenholtz (2004) and Kalai et al. (2010). Specifically, we say that the self-referential game G permits detection if for any code of conduct profile $r \in R$ where $r_i = r_j$ for all i, j, and for each player i, there exists a subset of signals $Y_j^i \subset Y_j$ for all players $j \neq i$ such that $\pi_j(Y_j^i|\tilde{r}_i, r_{-i}) = 1$ for any $\tilde{r}_i \neq r_i$ and $\pi_j(Y_j^i|r) = 0$. We view this detection notion to be plausible in small communities, and when players delegate their play either to an agent (with irreversible compensation schemes) or to programs (for example financial trading and proxies; see Section 4).

Next, we state our first main result:

Theorem 1 *If $v_i = u_i(\sigma) \geq \underline{u}_i$ for all players i with strategy profile $\sigma \in \mathsf{S}$ and G permits detection, then there exists an $r \in R$ such that $(v_1, \ldots, v_N)$ is a Nash equilibrium payoff of G.*

Proof Take any $\sigma \in \mathsf{S}$ such that for any i, $u_i(\sigma) \geq \underline{u}_i$. Consider the code of conduct $r_i \in R_0$ that prescribes

$$r_i^j(y_j) = \begin{cases} \sigma_j & \text{if } y_j \notin Y_j^k \quad \text{for all } k \in I, \\ \underline{\sigma}_j^k & \text{otherwise.} \end{cases}$$

If all players choose r_i, any player i would get $U_i(r) = u_i(\sigma)$. Contrary, if player i adheres to some $\tilde{r}_i$ so that $\tilde{r}_i(y_i) = \tilde{\sigma}_i$ for all $y_i \in Y_i$ and any $\tilde{\sigma}_i \in \mathsf{S}_i$; and $\tilde{r}_i^j(y_j) = r_i^j(y_j)$, he gets $U_i(\tilde{r}_i, r_{-i}) = \underline{u}_i$. It follows then that r is a Nash equilibrium of the self-referential game. □

This theorem is similar to the "benchmark theorem" in Levine and Pesendorfer (2007) with the difference that we consider asymmetry and more than two players. Notice that to obtain a full folk theorem, for example, for games that do not have Pareto efficient payoffs in pure strategies or pure minmax strategies, we do not need to use jointly controlled lotteries as Kalai et al. (2010) require because players do not condition directly on opponents' codes of conduct, and hence they can commit to randomize after receiving information about opponents' codes of conduct. Since we can accommodate correlated strategies by defining S_i as a subset of such strategies and assuming that players have access to a public randomization device, our folk theorem attains efficiency unlike program equilibria (Tennenholtz 2004).

3.1 Only some players informed

While it may be a reasonable approximation that players can detect whether or not some other players use the same code of conduct as themselves, in many settings it is reasonable to suppose that they can do so only for a small subset of other players who they observe closely. Depending on the context, a small group of members in a community might be able to identify who comply with the established convention but may require the help of other members to inflict a social punishment on deviators. Relaxing the observability assumption, we imposed above creates a difficulty when establishing a folk theorem because players needs to communicate what they observe about others' codes of conduct, and they must have an incentive to report truthfully. We find that the property the self-referential game must satisfy to prove a folk theorem is that each player's code of conduct is observed by at least two other players when they can publicly communicate. This observability assumption is weaker than the one imposed by Tennenholtz (2004) and Kalai et al. (2010).

We assume a cheap talk communication stage: after receiving their private signals $y_i \in Y_i$, players simultaneously make announcements $z_i \in Z_0$ that can be observed by everyone.[5] The set of possible announcements Z_0 is finite and common to all players. A profile of announcements is $z \in Z = \times_i Z_0$. Note that the identity of both the announcer and the subset of players who are thought to have deviated are crucial. Let $z_i^D \in Z_0$ be player i's announcement pointing that a subset of opponents $D \subseteq I$ have chosen a different code of conduct (that is some players potentially adhere to a different social norm). We require that there be at least 2^N of such possible announcements, that is, $\#Z_0 \geq 2^N$. We allow for not sending a message $\{\emptyset\} \in Z_0$.

A strategy for player i consists of an announcement policy $m_i : Y_i \to Z_0$ and an implementation policy $\phi_i(\cdot, y_i) : Z \to \mathsf{S}_i$ given any private signal $y_i \in Y_i$. In an extended self-referential game E, a code of conduct now specifies $r_i^j = (m_i^j, \phi_i^j)$ for each $j \in I$, which belongs to the common space of codes of conduct

$$R_1 = \Big\{ r_i = (m_i^j, \phi_i^j)_{j \in I} : \forall j \in I, m_i^j \in Z_0^{Y_j}, \phi_i^j \in \mathsf{S}_j^{Z \times Y_j} \text{ and } \forall y_j \in Y_j, m_i^j(y_j) \in Z_0, \forall (z, y_j) \in Z \times Y_j, \phi_i^j(z, y_j) \in \mathsf{S}_j \Big\}.$$

As before codes of conduct not only determine behavior as a function of signals, but also the probability distribution over the signals; for $r \in \times_i R_1$ we continue to denote this by $\pi(\cdot|r) \in \Delta(Y)$.

To prove a folk theorem, we require that each player is observed by at least two other players. This avoids the possibility that one player deviates, and at the same time, points the finger at the only person who is able to monitor her. In this case, the remaining players do not know who to punish and may not be able to punish both. Formally, we say that the extended self-referential game E weakly permits detection,

[5] Note that we do not require complicated protocols between players or the participation of a mediator as in Forges (1986) and Myerson (1986). Also, the sort of communication we model is akin to Forges (1990) as players informally talk to everyone and the order of how announcements are made does not change our results.

meaning that for any code of conduct profile $r \in R$ with $r_i = r_j$ for all i, j, and for any player i there is subsets of signals $Y_j^i \subset Y_j, Y_k^i \subset Y_k$ for at least two distinct players $j, k \neq i$ such that $\pi_j(Y_j^i|\tilde{r}^i, r^{-i}) = \pi_k(Y_k^i|\tilde{r}^i, r^{-i}) = 1$ for any $\tilde{r}_i \neq r_i$ and $\pi_j(Y_j^i|r) = \pi_k(Y_k^i|r) = 0$. What weak detection says in a sense is that there are "neutral" witnesses, that is, people who observe wrongdoing but who cannot be credibly accused of wrongdoing by the wrongdoer and this information is common knowledge in the society.

We are now in a position to state the second main result of the paper:

Theorem 2 *For all $\sigma \in \mathsf{S}$ such that $v_i = u_i(\sigma) \geq \underline{u}_i$ for all $i \in I$, if the extended self-referential game E weakly permits detection, then there is an $r \in R$ such that $(v_1, \ldots, v_N)$ is a Nash equilibrium payoff of E.*

Proof Take $\sigma \in \mathsf{S}$ such that $u_i(\sigma) \geq \underline{u}_i$ for all i. We construct $r_i \in R_0$ as follows. Let m_i^j be such that $m_i^j(y_j) = z_j^k$ if $y_j \in Y_j^k$ and $m_i^j(y_j) = \{\emptyset\}$ otherwise. Also, $\phi_i^j(z, y_j) = \underline{\sigma}_j^k$ for all $y_j \in Y_j^k$ or for all $z \in Z$ such that $z_l^k \in z$ for some $l, k \in I$; otherwise $\phi_i^j(z, y_j) = \sigma_j$. If all players choose r_i then $U_i(r) = u_i(\sigma)$. We begin by checking potential deviations. It suffices to check the following cases. Suppose player i chooses $\tilde{r}_i$ with $\tilde{m}_i^i(y_i) = z_i^k$ for all $y_i \in Y_i$ and $\tilde{\phi}_i^i(z) = \tilde{\sigma}_i \in \mathsf{S}_i$ for all $y_i \in Y_i$, $z \in Z$. By weak permit detection, there is a player $j \neq i$ that receives $y_j \in Y_j^i$ and would announce z_j^i so player i would face his minmax payoff $U_i(\tilde{r}_i, r_{-i}) = \underline{u}_i$, even if player $j = k$ there will be a mutually accusation which is not possible. Alternatively, player i chooses $\tilde{r}_i$ with $\tilde{m}_i^i(y_i) = m_i^i(y_i)$ and $\tilde{\phi}_i^i(z, y_i) = \tilde{\sigma}_i \in \mathsf{S}_i$ for all $y_i \in Y_i$, $z \in Z$. Again, by weak permit detection, there is a player $j \neq i$ that receives $y_j \in Y_j^i$ so will make the announcement z_j^i, and therefore player i would obtain his minmax payoff $U_i(\tilde{r}_i, r_{-i}) = \underline{u}_i$. □

The reason for requiring that each player is monitored by at least two other players is that if players respond to unique announcements, then a player can always foil the system by choosing a different code of conduct and announcing another player has "misbehaved," since communication is based on cheap talk messages. At worst, when he is detected there will be two such announcements so his opponents not only do not know who to punish, but also may not be able to punish both announcers. However, if the self-referential game weakly permits detection then we can specify that when three players announce deviations and one points to the others, then the player who has two accusations is punished. In addition to circumventing the problem of meaningful announcements, communication raises the issue that players may not report observations if they would be punished by doing so. But such incentives are also corrected within the self-referential framework since the cheap talk messages are part of the code of conduct, and consequently can be detected. This is why costless communication becomes a "signal" device in this context. Thus, we construct equilibria where players voluntarily communicate what they observed about their opponents' codes of conduct and show that communication is a powerful tool to allow players to coordinate behavior when they do not share consensus on whether all players comply with the conventional norm.

4 Application: codes of conduct as computer algorithms

As online markets grow, people are using more often computer programs to trade on their behalf, such as "proxies" that bid on online auctions and/or keep track of posted prices, and such as click stream pricing techniques used by many websites. Sellers that operate through websites have lots of commitment power and could use computer programs employed by the buyers. Therefore, a natural physical model of strategies is that players play by submitting computer programs to play on their behalf. We next describe how self-referential games formalize these ideas and encompass the notion of program equilibrium (Tennenholtz 2004).

In the self-referential framework, computer programs work as follows. Fix a signal profile set Y and break the program into two parts, one of which generates Y based on analyzing the programs, the other of which maps Y into the strategy profile set S. The programs also receive as input the program of the other player, that is, programs work as files as well. A well-known result is the impossibility of running an algorithm in which we are able to read the opponent's program and best respond to it. On the other hand, it is still possible to write down a computer program that makes a binary choice: give one response if both programs are the same, and give an alternative response if different. Specifically, there is a finite language L of computer statements, and a finite limit l on the length of a program.[6] The space of computer programs is $P = \{(x_n)_{n=1}t \in L|t \leq l\}$, the set of all sequences in L of length less than or equal to l. Each program $p_i \in P$ produces outputs $p_i : P \times P \to \{1, 2, \ldots, \infty\} \times \mathsf{S}_i$.[7] The interpretation is that the program $p_i(p_1, p_2) = (\nu_i, \sigma_i)$ produces the result σ_i after ν_i steps. In case $\nu_i = \infty$, the program does not halt. Notice that depending on L these programs can be either Turing or finite state machines. A self-referential strategy is a pair consisting of a default strategy $\overline{\sigma}_i \in \mathsf{S}_i$ and a program, $r_i = (\overline{\sigma}_i, p_i)$. After players submit their programs p_1, p_2, each program p_i is given itself and the program submitted by the opposing player p_{-i} as inputs. All programs are halted after $\overline{\nu}$ steps. If $p_i(p_1, p_2) = (\nu_i, \sigma_i)$ and $\nu_i \leq \overline{\nu}$, that is, the program halted in time we then define the mapping $r_i : P \times P \to \mathsf{S}_i$ as follows: $r_i(p_1, p_2) = \sigma_i$, otherwise $r_i(p_1, p_2) = \overline{\sigma}_i$. Finally, to map these computer algorithms to a self-referential game we take $Y = \mathsf{S}$. The probability distribution $\pi(\cdot|r)$ is given by $\pi(y|r) = 1$ if $y_i = r_i(p_1, p_2)$ for all i, and $\pi(y|r) = 0$ otherwise.

We now define a specific notion of self-referential games that is suitable for this context.

Definition 1 The strategy space S_i is *self-referential with respect to the deadline* $\overline{\nu}$ if for every pair of actions $\overline{a}_i, \underline{a}_i$ there exists a strategy $\sigma_i = (d_i, p_i) \in \mathsf{S}_i$, such that

$$p_i(\tilde{d}, \tilde{p}) = \begin{cases} \overline{\nu}, \overline{a}_i & \text{if } \tilde{d} = d_i, \tilde{p} = p_i, \\ \nu_i, \underline{a}_i & \text{otherwise.} \end{cases}$$

[6] For ease of exposition, we do not describe the formal metalanguage, but we assume a standard language that allows logic statements (see, for example, Hopcroft and Ullman 1979).

[7] This analysis can be readily extended to the case that there are more than two players.

The next example shows that there are games that satisfy this definition of self-referential strategies and illustrates that we can easily construct self-referential equilibria with efficient outcomes that are not feasible in the baseline game for any Nash equilibria.

Example 1 We consider a two-player trading game where each player owns an object that can be traded with the opponent. The action space is $A = \{0, 1\}$, where 0 represents "not trade with your opponent" and 1 represents "trade with your opponent." The good is worth $\gamma > 1$ to the opponent, and 1 to the owner. The players' dominant strategy is to keep the good for themselves and the Nash equilibrium of the game is $a = (0, 0)$.

We will show that instead of an entire strategy space, a simple strategy satisfies definition 1 and that the cooperative outcome can be sustained in a self-referential equilibrium. Let the default strategy be $\overline{\sigma}_i = 0$, no trading. Here, the language L is the Windows command language and the listing (program) p_i is given below:

```
@echo off
if "0" EQU "%3" goto sameactions
echo 0
goto finish
:sameactions
echo n | comp %2 %4
if %errorlevel% EQU 0 goto cooperate
echo 0
goto finish
:cooperate
echo 1
:finish
```

This program runs from the Windows command line, and takes as inputs four arguments: a digit describing the "own" default action, an "own" filename, an opponent default action and an opponent filename. If the opponent default action is 0 and the opponent program p_{-i} is identical to the listing above, the program p_i generates as its final output the number 1; otherwise, it generates the number 0.[8] Since it has access to sequence of its own instructions, it compares them to the sequence of opponent program instructions to check if they are the same or not.[9] Although in this listing all the actual work is done by the "comp" command it is easy enough to write a program that compares two files, and takes a number of steps proportional to the length of the shorter file. In other words, the program works in finite, and relatively short time.

When both players choose the above program, the two programs are syntactically the same, and then both choose $a = 1$ and obtain γ. Yet, if only one player submits a different program that differs syntactically from the proposed listing, that player receives at most $1 < \gamma$. We have shown that there is a self-referential equilibrium

[8] Note that we could allow for a mixed strategy, for example, when the program calls for action 1 instead it could call action 1 with some positive probability $1 > \alpha > 0$.

[9] In the same spirit of social norms, the fact that the program has access to the syntax sequence of the other program it does not imply that it can execute the other program's output.

for the trading game that yields always trading by both players if the output actions prescribed by each program are executed simultaneously. Notice that we might be able to sustain any feasible individually rational payoff, as we showed in Theorem 1, if we replace the default strategy by the minmax strategy and target strategy by the aimed individually rational strategy profile.

5 Conclusion

We showed a folk theorem for normal form games where players observe perfectly informative signals that point at deviant codes of conduct, and hence deviators are punished with certainty. We further weakened the assumption about who observe these signals, highlighting the importance of communication in situations where players have the ability to use conditional commitment devices. We provided an important application of codes of conduct in this specific environment and described how these strategies can be represented as computer algorithms.

Acknowledgements We thank Drew Fudenberg, Phillip Johnson, George Mailath, Wolfgang Pesendorfer, Balazs Szentes, and Satoru Takahashi for helpful discussions. Juan Block acknowledges support from the Cambridge-INET Institute, and David Levine thanks the National Science Foundation (Grant SES-0851315) and the European University Institute for financial support.

References

Bachi, B., Ghosh, S., Neeman, Z.: Communication and deception in 2-player games. Working paper (2014)

Block, J.I.: Timing and codes of conduct. Working paper (2013)

Block, J.I., Levine, D.K.: Codes of conduct, private information and repeated games. Int. J. Game Theory (2015). doi:10.1007/s00182-015-0498-2

Forges, F.: An approach to communication equilibria. Econometrica **54**(6), 1375–1385 (1986)

Forges, F.: Universal mechanisms. Econometrica **58**(6), 1341–1364 (1990)

Frank, R.H.: Passions Within Reason: The Strategic Role of the Emotions. W. W. Norton and Company, USA (1988)

Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages, and Computation. Addison-Wesley, USA (1979)

Kalai, A.T., Kalai, E., Lehrer, E., Samet, D.: A commitment folk theorem. Games Econ. Behav. **69**(1), 127–137 (2010)

Levine, D.K., Pesendorfer, W.: The evolution of cooperation through imitation. Games Econ. Behav. **58**(2), 293–315 (2007)

Myerson, R.B.: Multistage games with communication. Econometrica **54**(2), 323–358 (1986)

Schelling, T. C.: The Strategy of Conflict. Harvard University Press, USA (1960)

Tennenholtz, M.: Program equilibrium. Games Econ. Behav. **49**(2), 363–373 (2004)