Commitment Games with Conditional Information Disclosure

Anthony DiGiovanni[*], **Jesse Clifton**[*]

Center on Long-Term Risk, London, UK
{anthony.digiovanni, jesse.clifton}@longtermrisk.org

Abstract

The conditional commitment abilities of mutually transparent computer agents have been studied in previous work on commitment games and program equilibrium. This literature has shown how these abilities can help resolve Prisoner's Dilemmas and other failures of cooperation in complete information settings. But inefficiencies due to private information have been neglected thus far in this literature, despite the fact that these problems are pervasive and might also be addressed by greater mutual transparency. In this work, we introduce a framework for commitment games with a new kind of conditional commitment device, which agents can use to conditionally disclose private information. We prove a folk theorem for this setting that provides sufficient conditions for ex post efficiency, and thus represents a model of ideal cooperation between agents without a third-party mediator. Further, extending previous work on program equilibrium, we develop an implementation of conditional information disclosure. We show that this implementation forms program ϵ-Bayesian Nash equilibria corresponding to the Bayesian Nash equilibria of these commitment games.

Introduction

What are the upper limits on the ability of rational, self-interested agents to cooperate? As autonomous systems become increasingly responsible for important decisions, including in interactions with other agents, the study of "Cooperative AI" (Dafoe et al. 2020) will potentially help ensure these decisions result in cooperation. It is well-known that game-theoretically rational behavior — which will potentially be more descriptive of the decision-making of advanced computer agents than humans — can result in imperfect cooperation, in the sense of inefficient outcomes. Some famous examples are the Prisoner's Dilemma and the Myerson-Satterthwaite impossibility of efficient bargaining under incomplete information (Myerson and Satterthwaite 1983). Fearon (1995) explores "rationalist" explanations for war (i.e., situations in which war occurs in equilibrium); these include Prisoner's Dilemma-style inability to credibly commit to peaceful alternatives to war, as well as incentives to misrepresent private information (e.g., military strength). Because private information is so ubiquitous in real strategic interactions, resolving these cases of inefficiency is a fundamental open problem. Inefficiencies due to private information will be increasingly observed in machine learning, as machine learning is used to train agents in complex multi-agent environments featuring private information, such as negotiation. For example, Lewis et al. (2017) found that when an agent was trained with reinforcement learning on negotiations under incomplete information, it failed to reach an agreement with humans more frequently than a human-imitative model did.

But greater ability to make commitments and share private information can open up more efficient equilibria. Computer systems could be much better than humans at making their internal workings legible to other agents, and at making sophisticated conditional commitments. More mutually beneficial outcomes could also be facilitated by new technologies like smart contracts (Varian 2010). This makes the game theory of interactions between agents with these abilities important for the understanding of Cooperative AI — in particular, for developing an ideal standard of multi-agent decision making with future technologies. An extreme example of the power of greater transparency and commitment ability is Tennenholtz (2004)'s "program equilibrium" solution to the one-shot Prisoner's Dilemma. The players in Tennenholtz's "program game" version of the Prisoner's Dilemma submit computer programs to play on their behalf, which can condition their outputs on each other's source code. Then a pair of programs with the source code `''If counterpart's source code == my source code: Cooperate; Else: Defect''` form an equilibrium of mutual cooperation.

In this spirit, we are interested in exploring the kinds of cooperation that can be achieved by agents who are capable of extreme mutual transparency and credible commitment. We can think of this as giving an upper bound on the ability of advanced artificially intelligent agents, or humans equipped with advanced technology for commitment and transparency, to achieve efficient outcomes. While such abilities are inaccessible to current systems, identifying sufficient conditions for cooperation under private information provides directions for future research and development, in order to avoid failures of cooperation. These are our main contributions:

1. We develop a new class of games in which players can

[*] These authors contributed equally.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

condition both their commitments and disclosure of private information on their counterparts' commitments and decisions to disclose private information. We present a folk theorem for these games: The set of equilibrium payoffs equals the set of feasible and interim individually rational payoffs, notably including all ex post efficient payoffs. The equilibria are conceptually straightforward: For a given ex post payoff profile, players disclose their private information and play according to an action profile attaining that payoff profile; if anyone deviates, they revert to a punishment policy (without disclosing private information to the deviator). The problem is to avoid circularity in these conditional decisions. Our result builds on Forges (2013)' folk theorem for Bayesian games without conditional information disclosure, in which equilibrium payoffs must also be incentive compatible. This expansion of the set of equilibrium payoffs is important, because in several settings, such as those of the classic Myerson-Satterthwaite theorem (Myerson and Satterthwaite 1983), ex post efficiency (or optimality according to some function of social welfare) and incentive compatibility are mutually exclusive.

2. In these commitment games, the conditional commitment and disclosure devices are abstract objects. The devices in Forges (2013)' and our folk theorems avoid circularity by conditioning decisions on the particular identities of the other players' devices, but this precludes robust cooperation with other devices that would output the same decisions. Using computer programs as conditional commitment and disclosure devices, we give a specific implementation of ϵ-Bayesian Nash equilibria corresponding to the equilibria of our commitment game. This approach extends Oesterheld (2019)'s "robust program equilibria." We solve the additional problems of (1) ensuring that the programs terminate with more than two players, (2) in circumstances where cooperating with other players requires knowing their private information. Ours is the first study of program equilibrium (Tennenholtz 2004) under private information.

Related Work

Commitment games and program equilibrium. We build on *commitment games*, introduced by Kalai et al. (2010) and generalized to Bayesian games (without verifiable disclosure) by Forges (2013). In a commitment game, players submit commitment devices that can choose actions conditional on other players' devices. This leads to folk theorems: Players can choose commitment devices that conditionally commit to playing a target action (e.g., cooperating in a Prisoner's Dilemma), and punishing if their counterparts do not play accordingly (e.g., defecting in a Prisoner's Dilemma if counterparts' devices do not cooperate). A specific kind of commitment game is one played between computer agents who can condition their behavior on each other's source code. This is the focus of the literature on program equilibrium (Rubinstein 1998; Tennenholtz 2004; LaVictoire et al. 2014; Critch 2019; Oesterheld 2019; Oesterheld and Conitzer 2021). Peters and Szentes (2012) critique the program equilibrium framework as insufficiently robust to new contracts, because the programs in, e.g., Kalai et al. (2010)'s folk theorem only cooperate with the exact programs used in the equilibrium profile. Like ours, the commitment devices in Peters and Szentes (2012) can disclose their types and punish those that do not also disclose. However, their devices disclose unconditionally and thus leave the punishing player exploitable, restricting the equilibrium payoffs to a smaller set than that of Forges (2013) or ours.

Our folk theorem builds directly on Forges (2013). In Forges' setting, players lack the ability to disclose private information. Thus the equilibrium payoffs must be incentive compatible. We instead allow (conditional) verification of private information, which lets us drop Forges' incentive compatibility constraint on equilibrium payoffs. Our program equilibrium implementation extends Oesterheld (2019)'s robust program equilibrium to allow for conditional information disclosure.

Strategic information revelation. In games of strategic information revelation, players have the ability to verifiably disclose some or all of their private information. The question then becomes: How much private information should players disclose (if any), and how should other players update their beliefs based on players' refusal to disclose some information? A foundational result in this literature is that of *full unraveling*: Under a range of conditions, when players can verifiably disclose information, they will act as if all information has been disclosed (Milgrom 1981; Grossman 1981; Milgrom and Roberts 1986). This means the mere possibility of verifiable disclosure is often enough to avoid informational inefficiencies. However, there are cases where unraveling fails to hold and, even when verifiable disclosure is possible, informational inefficiencies persist and lead to welfare losses. This can be due to uncertainty about a player's ability to verifiably disclose (Dye 1985; Shin 1994) or disclosure being costly (Grossman and Hart 1980; Jovanovic 1982). But disclosure of private information can fail even without such uncertainty or costs (Kovenock, Morath, and Münster 2015; Martini 2018). We will show how these kinds of private information problems can be remedied with the commitment technologies of our framework (but not weaker ones, like those of Forges (2013)).

Preliminaries: Games of Incomplete Information and Inefficiency

Definitions

Let G be a Bayesian game with n players. Each player i has a space of types T_i, giving a joint type space $T = \bigtimes_{i=1}^{n} T_i$. At the start of the game, players' types are sampled by Nature according to the common prior q. Each player knows only their type. Player i's strategy is a choice of action $a_i \in \mathcal{A}_i$ for each type in T_i. Let $u_i(\mathbf{t}, \boldsymbol{a})$ denote player i's expected payoff in this game when the players have types $\mathbf{t} = (t_1, \dots, t_n)$ and follow an action profile $\boldsymbol{a} = (a_1, \dots, a_n)$. A Bayesian Nash equilibrium is an action profile $\boldsymbol{a}$ in which every player and type plays a best response with respect to the prior over other players' types: For all players i and all types t_i, $a_i(t_i) \in$

$\arg\max_{a_i' \in \mathcal{A}_i} \mathbb{E}_{\mathbf{t}_{-i} \sim q(\cdot|t_i)} u_i(\mathbf{t}, (a_i'(t_i), \boldsymbol{a}_{-i}(\mathbf{t}_{-i})))$. An ϵ-Bayesian Nash equilibrium is similar: Each player and type expects to gain at most ϵ (instead of 0) by deviating from $\boldsymbol{a}$.

We assume players can correlate their actions by conditioning on a trustworthy randomization signal $\mathcal{C}$.[1] For any correlated policy $\boldsymbol{\mu}$ (a distribution over action profiles), let $u_i(\mathbf{t}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\mu}} u_i(\mathbf{t}, \boldsymbol{a})$. When it is helpful, we will write $\boldsymbol{\mu}(\cdot|\mathbf{s})$ to clarify the subset of the type profile $\mathbf{s} \subseteq \mathbf{t}$ on which the correlated policy is conditioned. Let $(a_j, \boldsymbol{\mu}_{-j})$ denote a correlated policy such that player j plays a_j with probability 1, and the actions of players other than j are sampled from $\boldsymbol{\mu}_{-j}$ independently of a_j. Then, the following definitions will be key to our discussion:

Definition 1. *A payoff vector* $\mathbf{x}$ *as a function of type profiles is* ***feasible*** *if there is a correlated policy* $\boldsymbol{\mu}(\cdot|\mathbf{t})$ *such that, for all players* j *and types* $t_j \in T_j$, $x_j(t_j) = \mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)} u_j(\mathbf{t}, \boldsymbol{\mu})$.

Definition 2. *A payoff* $\mathbf{x}$ *is* ***interim individually rational (INTIR)*** *if, for each player* j*, there is a correlated policy* $\boldsymbol{\tau}_{-j}(\cdot|\mathbf{t}_{-j})$ *used by the other players such that, for all* $t_j \in T_j$, $x_j(t_j) \geq \max_{a_j \in \mathcal{A}_j} \mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)} u_j(\mathbf{t}, (a_j, \boldsymbol{\tau}_{-j}(\cdot|\mathbf{t}_{-j})))$.

The *minimax policy* $\boldsymbol{\tau}_{-j}$ is used by the other players to punish player j. The threat of such punishments will support the equilibria of our folk theorem. Players only have sufficient information to use this correlated policy if they disclose their types to each other. Moreover, the punishment can only work in general if they do *not* disclose their types to player j, because the definition of INTIR requires the deviating player to be uncertain about $\mathbf{t}_{-j}$. Since the inequalities hold for all $t_j \in T_j$, the players do not need to know player j's type to punish them.

Definition 3. *A feasible payoff* $\mathbf{x}$ *induced by* $\boldsymbol{\mu}$ *is* ***incentive compatible (IC)*** *if, for each player* j *and type pair* $t_j, s_j \in T_j$, $x_j(t_j) \geq \mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)} u_j((t_j, \mathbf{t}_{-j}), \boldsymbol{\mu}(\cdot|s_j, \mathbf{t}_{-j}))$.

Incentive compatibility means that, supposing players report their part of a type profile on which their correlated policy is conditioned, no player prefers to lie about their type.

Definition 4. *Given a type profile* $\mathbf{t}$*, a payoff* $\mathbf{x}$ *is* ***ex post efficient*** *(hereafter, "efficient") if there does not exist* $\tilde{\boldsymbol{\mu}}$ *such that* $u_i(\mathbf{t}, \tilde{\boldsymbol{\mu}}) \geq x_i(t_i)$ *for all* i *and* $u_{i'}(\mathbf{t}, \tilde{\boldsymbol{\mu}}) > x_{i'}(t_{i'})$ *for some* i'.

We will also consider games with strategic information revelation, i.e., Bayesian games where, immediately after learning their types and before playing $\boldsymbol{a}$, players can disclose their private information as follows (the "disclosure phase"). Players simultaneously each choose $\Theta_i = (\Theta_{ij})_{j \neq i}$, where each disclosure set Θ_{ij} is from some disclosure space $\mathcal{R}(t_i)$, a subset of $\mathcal{T}(t_i) = \{\Theta_i \subseteq T_i \mid t_i \in \Theta_i\}$. Then, each player j observes each Θ_{ij}, thus learning that player i's type is in Θ_{ij}, and conditions a_j on $(\Theta_{ij})_{i \neq j}$. As is standard in the framework of strategic information revelation, disclosure is *verifiable* in the sense that each Θ_{ij} must contain player i's true type; they cannot falsely "disclose" a different type. We will place our results on conditional type disclosure in the context of the literature on unraveling:

Definition 5. *Let* $\{\Theta_i\}_{i=1}^n$ *be the profile of disclosure set lists (as functions of types) in a Bayesian Nash equilibrium* σ *of a game with strategic information revelation. Then* σ *has* ***full unraveling*** *if* $\Theta_{ij}(t_i) = \{t_i\}$ *for all* i, j*, or* ***partial unraveling*** *if* $\Theta_{ij}(t_i)$ *is a strict subset of* T_i *for some* i, j.

Inefficiency: Motivating Example

Uncertainty about others' private information, and a lack of ability or incentive to disclose that information, can lead to inefficient outcomes in Bayesian Nash equilibrium (or an appropriate refinement thereof). Here is an example we use to illustrate how informational problems can be overcome under our assumptions, but not under the weaker assumption of unconditional disclosure ability.

Example 1 (War under incomplete information, adapted from Slantchev and Tarar (2011))**.** Two countries $i = 1, 2$ are on the verge of war over some territory. Country 1 offers a split of the territory giving fractions s and $1 - s$ to countries 1 and 2, respectively. If country 2 rejects this offer, they go to war. Each player wins with some probability (detailed below), and each pays a cost of fighting $c_i > 0$. The winner receives a payoff of 1, and the loser gets 0.

The countries' military strength determines the probability that country 2 wins the war, denoted $p(\theta)$. Country 1 doesn't know whether country 2's army is weak (with type θ^W) or strong (θ^S), while country 1's strength is common knowledge. Further, country 2 has a weak point, which country 1 believes is equally likely to be in one of two locations $v \in \{1, 2\}$. Thus country 2's type is given by $t_2 = \{\theta, v\}$. Country 1 can make a sneak attack on $\hat{v} \in \{1, 2\}$, independent of whether they go to war. Country 1 would gain z from attacking $\hat{v} = v$, costing $c_{A,2}$ for country 2. But incorrectly attacking $\hat{v} \neq v$ would cost $c_{A,1} > z$ for country 1, so country 1 would not risk an attack given a prior of $\frac{1}{2}$ on each of the locations. If country 2 discloses its full type by allowing inspectors from country 1 to assess its military strength θ, country 1 will also learn v.

If country 1 has a sufficiently low prior that country 2 is strong, then war occurs in the unique perfect Bayesian equilibrium when country 2 is strong. Moreover, this can happen even if the countries can fully disclose their private information to one another. In other words, the unraveling of private information does not occur, because player 2 will be made worse off if they allow player 1 to learn about their weak point (see Appendix A.1 in the extended version of this paper[2] for a formal argument). Thus, unconditional disclosure is not sufficient to allow efficiency in equilibrium in this example, motivating the use of the conditional disclosure devices defined in the next section.

Next, we formally introduce our framework for commitment games with conditional information disclosure and present our folk theorem.

[1] Even without a trusted third party that supplies a common correlation signal, players could choose to all condition on the same "natural" source of randomness.

[2] `http://arxiv.org/abs/2204.03484`

Commitment Games with Conditional Information Disclosure

Setup

Players are faced with a "base game" G, a Bayesian game with strategic information revelation as defined in Definitions. In our framework, a commitment game is a higher-level Bayesian game in which the type distribution is the same as that of G, and strategies are devices that define mappings from other players' devices to actions and disclosure in G (conditional on one's type). We assume $\{\{t_i\}, T_i\} \subseteq \mathcal{R}(t_i)$ for all players i and types t_i, i.e., players are at least able to disclose their exact types or not disclose any new information. They additionally have access to devices that can condition (i) their actions in G and (ii) the disclosure of their private information on other players' devices. Upon learning their type t_i, player i chooses a commitment device d_i from an abstract space of devices D_i. These devices are indices that, based on t_i, induce a *response function* and a *type disclosure function* (as detailed below). As in Kalai et al. (2010) and Forges (2013), we will define these functions so as to allow players to condition their decisions on each other's decisions without circularity.

Let C be the domain of the randomization signal $\mathcal{C}$ (a random variable), and $D_{-i} = \bigtimes_{j \neq i} D_j$. First, adopting the notation of Forges (2013), the response function is $r^i_{d_i(t_i)} : D_{-i} \times C \to \mathcal{A}_i$. Given the other players' devices $\mathbf{d}_{-i} = (d_j)_{j \neq i}$ and the realized value c of $\mathcal{C}$, player i's action in G after the disclosure phase is $r^i_{d_i(t_i)}(\mathbf{d}_{-i}, c)$.[3] Conditioning the response on c lets players commit to correlated distributions over actions.

Second, we introduce type disclosure functions $y^i_{d_i(t_i)} : D_{-i} \to \{0,1\}^{n-1}$, which are not in the framework of Forges (2013). The jth entry of $y^i_{d_i(t_i)}(\mathbf{d}_{-i})$ indicates whether player i discloses their type to player j, i.e., player j learns $\Theta_{ij} = \{t_i\}$ if this value is 1 or $\Theta_{ij} = T_i$ if it is 0. (We can restrict attention to cases where either all or no information is disclosed, as our folk theorem shows that such a disclosure space is sufficient to enforce each equilibrium payoff profile.) Thus, each player i can condition their action $r^i_{d_i(t_i)}(\mathbf{d}_{-i}, c)$ on the others' private information disclosed to them via $(y^j_{d_j(t_j)}(\mathbf{d}_{-j}))_{j \neq i}$. Further, they can choose whether to disclose their type to another player, via $y^i_{d_i(t_i)}(\mathbf{d}_{-i})$, based on that player's device. Thus players can decide not to disclose private information to players whose devices are not in a desired device profile, and instead punish them.

Then, the commitment game $G(\mathcal{D})$ is the one-shot Bayesian game in which each player i's strategy is a device $d_i \in D_i$, as a function of their type. After devices are simultaneously and independently submitted (potentially as a draw from a mixed strategy over devices), the value c is drawn from the randomization signal $\mathcal{C}$, and players play the induced action profile $(r^i_{d_i(t_i)}(\mathbf{d}_{-i}, c))_{i=1}^n$ in G. Thus the ex post payoff of player i in $G(\mathcal{D})$ from a device profile $\mathbf{d} = (d_i)_{i=1}^n$ is $u_i(\mathbf{t}, (r^i_{d_i(t_i)}(\mathbf{d}_{-i}, c))_{i=1}^n)$.

Folk Theorem

Our folk theorem consists of two results: First, any feasible and INTIR payoff can be achieved in equilibrium (Theorem 1). As a special case of interest, then, any efficient payoff can be attained in equilibrium. Second, all equilibrium payoffs in $G(\mathcal{D})$ are feasible and INTIR (Proposition 1). The proof of Proposition 1 is straightforward (see Appendix C).

Theorem 1. *Let $G(\mathcal{D})$ be any commitment game. For type profile $\mathbf{t}$, let $\boldsymbol{\mu}$ be a correlated policy inducing a feasible and INTIR payoff profile $(u_i(\mathbf{t}, \boldsymbol{\mu}))_{i=1}^n$. Let $\hat{\boldsymbol{\tau}}$ be a punishment policy that is arbitrary except, if j is the only player with $d'_j \neq d_j$, let $\hat{\boldsymbol{\tau}}$ be the minimax policy $\boldsymbol{\tau}_{-j}$ against player j. Conditional on the signal c, let $\boldsymbol{\mu}^c(\mathbf{t})$ be the deterministic action profile, called the target action profile, given by $\boldsymbol{\mu}(\cdot|\mathbf{t})$, and let $\hat{\boldsymbol{\tau}}^c$ be the deterministic action profile given by $\hat{\boldsymbol{\tau}}$. For all players i and types t_i, let d_i be such that:*

$$r^i_{d_i(t_i)}(\mathbf{d}'_{-i}, c) = \begin{cases} \mu^c_i(\mathbf{t}), & \textit{if } \mathbf{d}'_{-i} = \mathbf{d}_{-i} \\ \hat{\tau}^c_i, & \textit{otherwise,} \end{cases}$$

$$y^i_{d_i(t_i)}(\mathbf{d}'_{-i})_j = \begin{cases} 1, & \textit{if } d'_j = d_j \\ 0, & \textit{otherwise.} \end{cases}$$

Then, the device profile $\mathbf{d}$ is a Bayesian Nash equilibrium of $G(\mathcal{D})$.

Proof. We first need to check that the response and type disclosure functions only condition on information available to the players. If all players use $\mathbf{d}$, then by construction of $y^i_{d_i(t_i)}$ they all disclose their types to each other, and so are able to play $\boldsymbol{\mu}(\cdot|\mathbf{t})$ conditioned on their type profile (regardless of whether the induced payoff is IC). If at least one player uses some other device, the players who do use $\mathbf{d}$ still share their types with each other, thus they can play $\hat{\boldsymbol{\tau}}$.

Suppose player j deviates from $\mathbf{d}$. That is, player j's strategy in $G(\mathcal{D})$ is $d'_j \neq d_j$. Note that the outputs of player j's response and type disclosure functions induced by d'_j may in general be the same as those returned by d_j. We will show that $\hat{\boldsymbol{\tau}}^c$ punishes deviations from the target action profile regardless of these outputs, as long as there is a deviation in *functions* r'^j or y'^j. Let $a'_j = r^j_{d'_j(t_j)}(\mathbf{d}_{-j}, c)$. Then:

$$\begin{aligned} &\mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)}(u_j(\mathbf{t}, (a'_j, \hat{\boldsymbol{\tau}}))|d'_j, \mathbf{d}_{-j}) \\ &\quad = \mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)} u_j(\mathbf{t}, (a'_j, \boldsymbol{\tau}_{-j}(\cdot|\mathbf{t}_{-j}))) \\ &\quad \leq \mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)} u_j(\mathbf{t}, \boldsymbol{\mu}) \qquad \text{(by INTIR)} \\ &\quad = \mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)}(u_j(\mathbf{t}, \boldsymbol{\mu})|d_j, \mathbf{d}_{-j}). \end{aligned}$$

This last expression is the ex interim payoff of the proposed commitment d_j given that the other players use $\mathbf{d}_{-j}$, therefore $\mathbf{d}$ is a Bayesian Nash Equilibrium. □

Proposition 1. *Let $G(\mathcal{D})$ be any commitment game. If a device profile $\mathbf{d}$ is a Bayesian Nash equilibrium of $G(\mathcal{D})$, then the induced payoff $\mathbf{x}$ is feasible and INTIR.*

[3] A player who chooses to "not commit" submits a device that is not a function of the other players' devices. In this case, the other devices can only condition on this non-commitment choice, not on the particular action this player chooses.

Our assumptions do not imply the equilibrium payoffs are IC (unlike Forges (2013)). Suppose a player i's payoff would increase if the players conditioned the correlated policy on a different type (i.e., not IC). This does not imply that a profit is possible by deviating from the equilibrium, because in our setting the other players' actions are conditioned on the type disclosed by i. In particular, as in our proposed device profile, they may choose to play their part of the target action profile only if all other players' devices disclose their (true) types.

The assumptions that give rise to this class of commitment games with conditional information disclosure are stronger than the ability to unconditionally disclose private information. Recalling the unraveling results from Related Work, unconditional disclosure ability is sometimes sufficient for the full disclosure of private information, or for disclosure of the information that prohibits incentive compatibility, and thus the possibility of efficiency in equilibrium. But this is not always true, whereas efficiency is always attainable in equilibrium under our assumptions. In Appendix A, we first show that full unraveling fails in our motivating example when country 2 has a weak point. Then, we discuss conditions under which the ability to partially disclose private information is sufficient for efficiency, and examples where these conditions don't hold.

Implementation of Conditional Type Disclosure via Robust Program Equilibrium

Having shown that all efficient payoff profiles are achievable in equilibrium using conditional commitment and disclosure devices, we next consider how players can practically (and more robustly) implement these abstract devices. In particular, can players achieve efficient equilibria without using the exact device profile in Theorem 1, which can only cooperate with itself? We now develop an implementation showing that this is possible, after providing some background.

Oesterheld (2019) considers two computer programs playing a game. Each program can simulate the other in order to choose an action in the game. He constructs a program equilibrium — a pair of programs that form an equilibrium of this game — using "instantaneous tit-for-tat" strategies. In the Prisoner's Dilemma, the pseudocode for these programs (called "ϵGroundedFairBot") is: "With small probability ϵ: Cooperate; Else: do what my counterpart does when playing against me." These programs cooperate with each other and punish defection. Note that these programs are recursive, but guaranteed to terminate because of the ϵ probability that a program outputs Cooperate unconditionally.

We use this idea to implement conditional commitment and disclosure devices. For us, "disclosing private information and playing according to the target action profile" is analogous to cooperation in the construction of ϵGroundedFairBot. Thus, instead of a particular device profile in which a device cooperates if and only if all other devices are in that profile, we consider programs that cooperate if and only if all other programs output cooperation against each other. We will first describe the appropriate class of programs for *program games* under private information. Then we develop our program, ϵGroundedFairSIRBot (where "SIR" stands for "strategic information revelation"), and show that it forms a δ-Bayesian Nash equilibrium of a program game. Pseudocode for ϵGroundedFairSIRBot is given in Algorithm 1.

As in Setup, there is a base game G, and players choose strategies that implement actions in G conditional on each other's strategies. In a program game, programs fill the role of devices in a commitment game. Player i's strategy in the program game is a choice p_i from the program space P_i, a set of computable functions from $\bigtimes_{j=1}^{n} P_j \times C \times \{0,1\}$ to $\mathcal{A}_i \cup \{0,1\}^{n-1}$. A program returns either an action or a type disclosure vector (just as a device induces response and type disclosure functions, which return actions and disclosure vectors, respectively). Each program takes as input the players' program profile, the signal c, and a boolean that equals 1 if the program's output is an action, and 0 otherwise. For brevity, we write p_i^r for a call to a program with the boolean set to 1, otherwise p_i^y. Letting $\mathbf{p} = (p_m)_{m=1}^{n}$ be the players' program profile, player i's action in G is a call to their program $p_i(\mathbf{p}, c, 1)$. (We refer to these initial program calls as the *base calls* to distinguish them from calls made by other programs.) Then, the ex post payoff of player i in the program game is $u_i(\mathbf{t}, (p_j(\mathbf{p}, c, 1))_{j=1}^{n})$.

Like Oesterheld (2019), we will use programs that unconditionally terminate with some small probability. To generalize this idea to a setting with private information and more than two players, we now introduce some additional elements of the program game and our program. First, in addition to $\mathcal{C}$ in the base game, there is a randomization signal $\mathcal{C}^P$ on which programs can condition their outputs. By using $\mathcal{C}^P$ to correlate decisions to unconditionally terminate, our program profile will be able to terminate with probability 1, despite the exponentially increasing number of recursive program calls. In particular, $\mathcal{C}^P$ reads the call stack of the players' program profile. At each depth level L of recursion reached in the call stack, a variable U_L is independently sampled from Unif$[0,1]$. Each program call at level L can read off the values of U_L and U_{L+1} from $\mathcal{C}^P$. The index L itself is not revealed, however, because programs that "know" they are being simulated could defect in the base calls, while cooperating in simulations to deceive the other programs. Second, let ϵGroundedFairSIRBot^r and ϵGroundedFairSIRBot^y be calls to the program ϵGroundedFairSIRBot with output_action $= 1$ and output_action $= 0$, respectively. To ensure that our programs terminate in play with a deviating program, ϵGroundedFairSIRBot^r will call truncated versions of its counterparts' disclosure programs: For $p_i \in P_i$, let $[p_i]$ denote p_i with immediate termination upon calling another program.

Figure 1 visually summarizes a program game between ϵGroundedFairSIRBot and some other program. Like a device in the profile in Theorem 1, which checks if the other devices are part of a profile that disclose their types and play their parts of the target action profile ("cooperate"), our program checks if the other programs disclose and cooperate with it. With high probability, ϵGroundedFairSIRBot^r checks if all other players' programs disclose their types

Algorithm 1: ϵGroundedFairSIRBot

```
Require: Program profile $\mathbf{p}$, randomization signal value $c$,
    boolean output_action
 1: if output_action $= 1$ then
 2:     if $U_{L+1} \geq \epsilon$ then
 3:         for $k \neq i$ do    ▷ Check if each player discloses
 4:             $\mathbf{y}^k \leftarrow p_k(\mathbf{p}, c, 0)$
 5:         if $\mathbf{y}^k = \mathbf{1}$ for all $k \neq i$ then
 6:             if $U_L < \epsilon$ then ▷ Unconditionally cooperate
 7:                 return $\mu_i^c(\mathbf{t})$
 8:             for $k \neq i$ do    ▷ Check if others cooperate
 9:                 $a_k \leftarrow p_k(\mathbf{p}, c, 1)$
10:             if $a_k = \mu_k^c(\mathbf{t})$ for all $k \neq i$ then
11:                 return $\mu_i^c(\mathbf{t})$
12:         return $\hat{\tau}_i^c$    ▷ Punish given known $(\mathbf{y}^m)_{m \neq i}$
13:     for $k \neq i$ do    ▷ Check truncated $p_k^y$
14:         $\mathbf{y}^k \leftarrow [p_k](\mathbf{p}, c, 0)$
15:     if $\mathbf{y}^k = \mathbf{1}$ for all $k \neq i$ then    ▷ Full type known
16:         return $\mu_i^c(\mathbf{t})$
17:     return $\hat{\tau}_i^c$
18: else
19:     $\mathbf{y}^i \leftarrow \mathbf{0}$
20:     if $U_L < \epsilon$ then    ▷ Unconditionally disclose
21:         return $\mathbf{1}$
22:     for $k \neq i$ do
23:         $\mathbf{y}^k \leftarrow p_k(\mathbf{p}, c, 0)$
24:         $a_k \leftarrow p_k(\mathbf{p}, c, 1)$
25:     if $\mathbf{y}^k = \mathbf{1}$ and $a_k = \mu_k^c(\mathbf{t})$ for all $k \neq i$ then
26:         return $\mathbf{1}$
27:     for $k \neq i$ do
28:         if $\mathbf{y}_i^k = 1$ and ($a_k = \hat{\tau}_k^c$ or $U_{L+1} < \epsilon$) then
29:             $\mathbf{y}_k^i \leftarrow 1$
30:     return $\mathbf{y}^i$
```

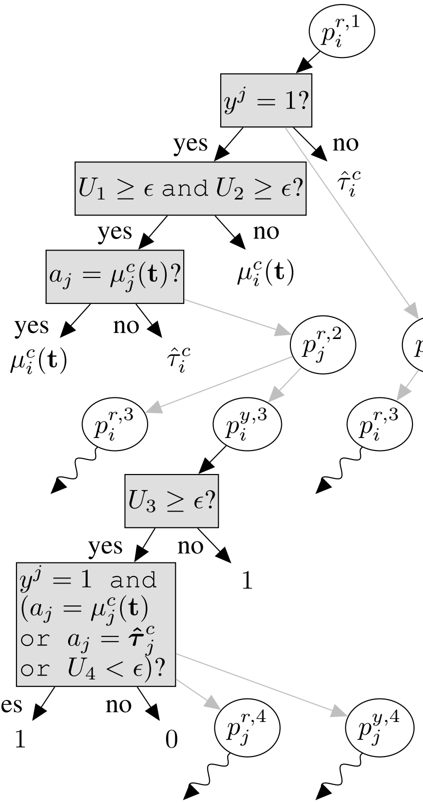

Figure 1: Flowchart for a 2-player program game between player i using ϵGroundedFairSIRBot, and player j using an arbitrary program. An edge to a white node indicates a call to the program in that node; to a gray node indicates a check of the condition in that node; and to a node without a border indicates the output of the most recent parent white node. Wavy edges depict a call to the program in the parent node, with its child nodes omitted for space. Superscripts indicate the level of recursion.

(lines 2-5 of Algorithm 1). If so, either with a small probability it unconditionally cooperates (lines 6-7), or it cooperates only when all other programs cooperate (lines 8-11). Otherwise, it punishes (line 12). If, with low probability, the next call to ϵGroundedFairSIRBot^r will unconditionally cooperate, then the current call cooperates if and only if the other *truncated* programs disclose (lines 13-17).

In turn, ϵGroundedFairSIRBot^y discloses its type unconditionally with probability ϵ (lines 20-21). Otherwise, it discloses to a given player j under two conditions (lines 25 and 28). First, player j must disclose to the user. Second, they must play an action consistent with the desired equilibrium, i.e., cooperate when all players disclose their types, or punish otherwise.

Unconditionally disclosing one's type and playing the target action avoids an infinite regress. Crucially, these unconditional cooperation outputs are correlated via $\mathcal{C}^P$. Therefore, in a profile of copies of this program, either all copies unconditionally cooperate together, or none of them do so. Using this property, we can show (see proof of Theorem 2 in Appendix D) that a profile where all players use this program outputs the target action profile with certainty. If one player deviates, first, ϵGroundedFairSIRBot^r immediately punishes if that player does not disclose. If the deviating player does disclose, with some small probability the other players unconditionally cooperate (lines 13-16), making this strategy slightly exploitable, but otherwise the deviator is punished. Even if a deviation is punished, ϵGroundedFairSIRBot^y may unconditionally disclose. In our approach, this margin of exploitability is the price of implementing conditional commitment and disclosure with programs that cooperate based on counterparts' outputs, rather than a strict matching of devices, without an infinite loop. Further, since a player is only able to unconditionally cooperate under incomplete information if they know all players' types, ϵGroundedFairSIRBot^r needs to prematurely terminate calls to programs that don't immediately unconditionally cooperate, but which may otherwise cause infinite recursion (line 14). This comes at the expense of robustness: ϵGroundedFairSIRBot punishes some players who may have otherwise cooperated, with low probability.

Theorem 2. *Consider the program game induced by a base game G and the program spaces $\{P_i\}_{i=1}^n$. Assume all strategies returned by these programs are computable. For a type profile $\mathbf{t}$, let $\boldsymbol{\mu}(\cdot|\mathbf{t})$ induce a feasible and INTIR payoff profile $(u_i(\mathbf{t}, \boldsymbol{\mu}))_{i=1}^n$. Let $\hat{\boldsymbol{\tau}}$ be the minimax policy if one player j deviates, and arbitrary otherwise.*

Let $\overline{u}$ be the maximum payoff achievable by any player in G, and $\delta = \overline{u}((1-\epsilon)^{-2} - 1)$. Then the program profile $\mathbf{p}$ given by Algorithm 1 (with `output_action` $= 1$) for players $i = 1, \dots, n$ is a δ-Bayesian Nash equilibrium. That is, if players $i \neq j$ play this profile, and player j plays a program $p'_j \in P_j$ that terminates with probability 1 given that any programs it calls terminate with probability 1, then:

$$\begin{aligned}&\mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)}(u_j(\mathbf{t}, \boldsymbol{\mu})|p'_j, \mathbf{p}_{-j}) \\ &\quad \leq \delta + \mathbb{E}_{\mathbf{t}_{-j} \sim q(\cdot|t_j)}(u_j(\mathbf{t}, \boldsymbol{\mu})|p_j, \mathbf{p}_{-j}).\end{aligned}$$

PROOF SKETCH. We need to check (1) that the program profile $\mathbf{p}$ terminates (a) with or (b) without a deviation, (2) that everyone plays the target action profile when no one deviates, and (3) that with high probability a deviation is punished. First, suppose no one deviates. If $U_L < \epsilon$ for two levels of recursion in a row, the calls to p_i^y and p_i^r all unconditionally disclose (line 21) and output the target action (line 16), respectively. Because these unconditional cooperative outputs are correlated through $\mathcal{C}^P$, the probability that $U_L < \epsilon$ at each pair of subsequent levels in the call stack is a nonzero constant. Thus it is guaranteed to occur eventually and cause termination in finite time, satisfying (1b). Moreover, each call to p_i^y or p_i^r in previous levels of the stack sees that the next level cooperates, and thus cooperates as well, ensuring that the base calls all output the target action profile. This shows (2).

If, however, one player deviates, we use the same guarantee of a run of subsequent $U_L < \epsilon$ events to guarantee termination. First, all calls to non-deviating programs terminate, because any call to ϵ`GroundedFairSIRBot`^{r} conditional on $U_{L+1} < \epsilon$ forces termination (line 14) of calls to other players' disclosure programs. Thus the deviating programs also terminate, since they call terminating non-deviating programs. This establishes (1a). Finally, in the high-probability event that the first two levels of calls to $\mathbf{p}$ do *not* unconditionally cooperate, ϵ`GroundedFairSIRBot`^{r} punishes the deviator as long as they do not disclose their type and play their target action. The punishing players will know each other's types, since a call to ϵ`GroundedFairSIRBot`^{y} is guaranteed by line 28 to disclose to anyone who also punishes or *unconditionally* cooperates in the next level. Condition (3) follows. □

We now discuss two practical considerations for this program equilibrium implementation. First, one obstacle to this implementation is demonstrating to one's counterpart that one's behavior is actually governed by the source code that has been shared. In our program game with private information, there is the additional problem that, as soon as one's source code is shared, one's counterpart may be able to read off one's private information (without disclosing their own). Addressing this in practice might involve modular architectures, where players could expose the code governing their strategy without exposing the code for their private information. Alternatively, consider AI agents that can place copies of themselves in a secure box, where the copies can inspect each other's full code but cannot take any actions outside the box. These copies read each other's commitment devices off of their source code, and report the action and type outputs of these devices to the original agents. If any copy within the box attempts to transmit information that another agent's device refused to disclose, the box deletes its contents. This protocol does not require a mediator or arbitrator; the agents and their copies make all the relevant strategic decisions, with the box only serving as a security mechanism. Applications of secure multi-party computation to machine learning (Knott et al. 2021), or privacy-preserving smart contracts (Kosba et al. 2016) — with the original agents treated as the "public" from whom code shared among the copies is kept private — might facilitate the implementation of our proposed commitment devices.

Second, it is an open question how to implement ϵ`GroundedFairSIRBot` in machine learning. We believe that this algorithm can be implemented with neural networks, by substituting in learned strategies for the hard-coded parts of the ϵ`GroundedFairSIRBot` algorithm that output decisions (lines 7, 11, 16). Indeed, Hutter (2021) takes this approach to applying multi-agent reinforcement learning to programs in the class of ϵ`GroundedFairBot`, of which ϵ`GroundedFairSIRBot` is a generalization.

Discussion

We have defined a new class of commitment games that allow disclosure of private information conditioned on other players' commitments. Our folk theorem shows that in these games, efficient payoffs are always attainable in equilibrium, which is not true in general without conditional disclosure devices. Finally, we have provided an implementation of this framework via robust program equilibrium, which can be used by computer programs that read each other's source code.

While conceptually simple, satisfying these assumptions in practice requires a strong degree of mutual transparency and conditional commitment ability, which is not possessed by contemporary human institutions or AI systems. Thus, our framework represents an idealized standard for bargaining in the absence of a trusted third party, suggesting research priorities for the field of Cooperative AI (Dafoe et al. 2020). The motivation for work on this standard is that AI agents with increasing economic capabilities, which would exemplify game-theoretic rationality to a stronger degree than humans, may be deployed in contexts where they make strategic decisions on behalf of human principals (Geist and Lohn 2018). Given the potential for game-theoretically rational behavior to cause cooperation failures (Myerson and Satterthwaite 1983; Fearon 1995), it is important that such agents are developed in ways that ensure they are able to cooperate effectively.

Commitment devices of this form would be particularly useful in cases where centralized institutions (Dafoe et al.

(2020), Section 4.4) for enforcing or incentivizing cooperation fail, or have not been constructed due to collective action problems. This is because our devices do not require a trusted third party, aside from correlation signals. A potential obstacle to the use of these commitment devices is lack of coordination in development of AI systems. This may lead to incompatibilities in commitment device implementation, such that one agent cannot confidently verify that another's device meets its conditions for trustworthiness and hence type disclosure. Given that commitments may be implicit in complex parametrizations of neural networks, it is not clear that independently trained agents will be able to understand each other's commitments without deliberate coordination between developers. Our program equilibrium approach allows for the relaxation of the coordination requirements needed to implement conditional information disclosure and commitment. Coordination on target action profiles for commitment devices or flexibility in selection of such profiles, in interactions with multiple efficient and arguably "fair" profiles (Stastny et al. 2021), will also be important for avoiding cooperation failures due to equilibrium selection problems.

Acknowledgments

We thank Lewis Hammond for helpful comments on this paper, and thank Caspar Oesterheld both for useful comments and for identifying an important error in an earlier version of one of our proofs.

References

Critch, A. 2019. A parametric, resource-bounded generalization of Löb's theorem, and a robust cooperation criterion for open-source game theory. *The Journal of Symbolic Logic*, 84(4): 1368–1381.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2020. Open Problems in Cooperative AI. arXiv:2012.08630.

Dye, R. A. 1985. Disclosure of nonproprietary information. *Journal of accounting research*, 123–145.

Fearon, J. D. 1995. Rationalist explanations for war. *International organization*, 49(3): 379–414.

Forges, F. 2013. A folk theorem for Bayesian games with commitment. *Games and Economic Behavior*, 78: 64–71.

Geist, E.; and Lohn, A. J. 2018. How might artificial intelligence affect the risk of nuclear war? Rand Corporation.

Grossman, S. J. 1981. The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3): 461–483.

Grossman, S. J.; and Hart, O. D. 1980. Disclosure laws and takeover bids. *The Journal of Finance*, 35(2): 323–334.

Hutter, A. 2021. Learning in two-player games between transparent opponents. arXiv:2012.02671.

Jovanovic, B. 1982. Truthful disclosure of information. *The Bell Journal of Economics*, 36–44.

Kalai, A. T.; Kalai, E.; Lehrer, E.; and Samet, D. 2010. A commitment folk theorem. *Games and Economic Behavior*, 69(1): 127–137.

Knott, B.; Venkataraman, S.; Hannun, A.; Sengupta, S.; Ibrahim, M.; and van der Maaten, L. 2021. CrypTen: Secure Multi-Party Computation Meets Machine Learning. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Kosba, A.; Miller, A.; Shi, E.; Wen, Z.; and Papamanthou, C. 2016. Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts. In *2016 IEEE Symposium on Security and Privacy (SP)*, 839–858.

Kovenock, D.; Morath, F.; and Münster, J. 2015. Information sharing in contests. *Journal of Economics & Management Strategy*, 24: 570–596.

LaVictoire, P.; Fallenstein, B.; Yudkowsky, E.; Barasz, M.; Christiano, P.; and Herreshoff, M. 2014. Program Equilibrium in the Prisoner's Dilemma via Löb's Theorem. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Lewis, M.; Yarats, D.; Dauphin, Y. N.; Parikh, D.; and Batra, D. 2017. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. arXiv:1706.05125.

Martini, G. 2018. Multidimensional Disclosure. http://www.giorgiomartini.com/papers/multidimensional_disclosure.pdf. Accessed: 2023-03-17.

Milgrom, P.; and Roberts, J. 1986. Relying on the information of interested parties. *The RAND Journal of Economics*, 18–32.

Milgrom, P. R. 1981. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 380–391.

Myerson, R. B.; and Satterthwaite, M. A. 1983. Efficient mechanisms for bilateral trading. *Journal of economic theory*, 29(2): 265–281.

Oesterheld, C. 2019. Robust program equilibrium. *Theory and Decision*, 86(1): 143–159.

Oesterheld, C.; and Conitzer, V. 2021. Safe Pareto Improvements for Delegated Game Playing. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 983–991.

Peters, M.; and Szentes, B. 2012. Definable and Contractible Contracts. *Econometrica*, 80: 363–411.

Rubinstein, A. 1998. *Modeling Bounded Rationality*. The MIT Press.

Shin, H. S. 1994. The burden of proof in a game of persuasion. *Journal of Economic Theory*, 64(1): 253–264.

Slantchev, B. L.; and Tarar, A. 2011. Mutual optimism as a rationalist explanation of war. *American Journal of Political Science*, 55(1): 135–148.

Stastny, J.; Riché, M.; Lyzhov, A.; Treutlein, J.; Dafoe, A.; and Clifton, J. 2021. Normative Disagreement as a Challenge for Cooperative AI. arXiv:2111.13872.

Tennenholtz, M. 2004. Program equilibrium. *Games and Economic Behavior*, 49(2): 363–373.

Varian, H. R. 2010. Computer mediated transactions. *American Economic Review*, 100(2): 1–10.