

Privacy in Implementation*

Ronen Gradwohl[†]

Abstract

In most implementation frameworks, agents care only about the outcome and not at all about the way in which it was obtained. Additionally, typical mechanisms for full implementation involve the complete revelation of all private information. In this paper I consider the problem of full implementation with agents who may prefer to protect their privacy. I show that privacy-protecting implementation, while typically impossible with normal-form mechanisms, is achievable with extensive-form mechanisms.

Keywords: Nash implementation subgame perfect implementation privacy

JEL Classification Numbers: D82, C72

*I gratefully acknowledge NSF award #1216006. I would like to thank Ehud Kalai, Ariel Rubinstein, Yuval Salant, Ron Siegel, and Rakesh Vohra for helpful conversations about this research. I am also grateful to seminar participants at Northwestern University, Tel Aviv University, Hebrew University, Harvard, and MIT, as well as several anonymous reviewers, for valuable feedback.

[†]Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA. Email: r-gradwohl@kellogg.northwestern.edu.

1 Introduction

While decision makers and committee members base decisions on anticipated material outcomes, they are also often motivated by what their actions reveal about their private information. For example, a cabinet member who casts a vote in a cabinet meeting may care about the actual effect of his vote, but if the position he prefers is perceived as unfavorable by some of his constituents, he may be hesitant to expose his preference. Similarly, a board member may be concerned about the information his actions reveal to shareholders, and a faculty member participating in a faculty meeting may be concerned about the information revealed to the department chair.

The core of this paper is a model in which agents have such *information-sensitive* preferences—they care not only about the outcomes of an interaction, but also about the type and amount of private information that is revealed. There are many reasons agents might care about such information revelation: for example, the information that is potentially revealed may have material consequences in later interactions, but it may be difficult to model all future interactions and determine how they are affected by this information. Additionally, in some situations a planner may only have control over a particular interaction, even if the information that is revealed in that interaction may have material consequences for the agents in the future. So when a planner designs a mechanism for a particular interaction, he views the agents as having information-sensitive preferences. Alternatively, some information may be “embarrassing” to the agents, or they may simply not wish outsiders to be privy to their desires. In any case, such information-sensitive preferences cannot be captured by standard models in economics or game theory.

The goal of this paper is twofold. First, it highlights the subtleties involved in modeling and reasoning about such information-sensitive preferences. And second, it initiates a study of implementation for agents with such preferences and contrasts it with implementation in the standard setting.

More specifically, in this paper I study the strategic effects of information-sensitive preferences in the context of full implementation¹ with complete information. With

¹With full implementation, the desideratum is a mechanism in which *all* equilibria lead to socially optimal outcomes. For various surveys of this vast literature, see Jackson (2001), Maskin and Sjöström (2002), or Palfrey (2002).

complete information, agents already know each other’s private information. This assumption is made for simplicity,² but it is also plausible, for example, for cabinet members, board members, or faculty members who have participated in past joint interactions and have communicated with one another, and are thus aware of each other’s preferences. Since agents already know each other’s preferences, their privacy concerns are vis-à-vis some outside observers. These include, for example, the cabinet members’ constituents, the board members’ shareholders and CEO, and the faculty members’ department chair and dean. I also assume that the interaction is transparent to these outside observers—namely, that all communication that is part of the mechanism is visible to outside observers. This corresponds to publicizing the minutes of a cabinet meeting (as is common for some cabinets and prescribed by various Open Meetings Acts), making the minutes of a board meeting available to shareholders (as is required by law of boards of public companies), or having the CEO or department chair sit in on a board or faculty meeting.

In this framework, the presence of information-sensitive preferences introduces two issues into the theory of implementation: The first is that known mechanisms for full implementation may no longer work, since the set of equilibria under information-sensitive preferences may be different from the set of equilibria under standard preferences. The second issue arises from the observation that in most known mechanisms for full implementation, all private information is revealed.³ But if agents prefer to keep their information private, it may be desirable to design mechanisms that preserve privacy to some extent.

In this paper I examine both of the following questions: Are there mechanisms that are *information-sensitive* implementations, in which all equilibria with respect to information-sensitive preferences achieve a social optimum? And are there mechanisms that are *privacy-protecting*, in which some equilibria reveal none of the agents’ private information other than what is implied by the outcome itself?

Observe that the first question is not specifically about privacy, but rather about information-sensitive preferences more generally. In particular, this question and the

²In general, implementation theory solves two parallel problems: information aggregation and coordination amongst agents. Complete information simplifies by narrowing the focus to coordination. See, e.g., Gradwohl (2017) for a study of the effect of privacy concerns on information aggregation.

³See, in particular, Maskin (1999) and Moore and Repullo (1988).

answers that will follow apply also in settings where agents *want* to disclose some or all of their private information. The second question is specific to privacy, and an assumption that will underlie some of the answers is that agents prefer privacy. However, the possibility results for this question can be extended to a setting in which agents have more nuanced preferences over what information is revealed and, in particular, to a setting in which they want to reveal some or all of their private information.⁴

A Simple Example To illustrate the difficulties of implementation with information-sensitive preferences, as well as provide an overview of the results of this paper, I describe a simple example with three agents and two alternatives $\{0, 1\}$. Each agent can be one of two types $\{t_0, t_1\}$, where type t_i *intrinsically* prefers outcome i to outcome $1 - i$. This means that type t_i prefers outcome i to $1 - i$ when the information known to outsiders about his (and others') types is fixed. Furthermore, suppose that agents' information-sensitive preferences are such that, for any fixed outcome, they strictly prefer that their true type not be revealed, regardless of the information revealed about other players' types. That is, they prefer both privacy, in which outside observers do not learn their true type, and deception, in which outside observers incorrectly learn their type, to the true revelation of their type. I will later specify the agents' information-sensitive preferences when the outcome is not fixed.

Suppose a planner wishes to implement the majority of agents' intrinsic preferences. He may design a simple voting mechanism: each agent reports an action that can be 0 or 1, and the outcome is the majority of the actions. Suppose also that the planner expects agents to vote truthfully—they vote for i if their type is t_i . Now, voting truthfully is an equilibrium with standard preferences, but is it an equilibrium with information-sensitive preferences?

When preferences are information sensitive, equilibria relate to both the outcomes and revealed information as follows. For each profile of types, each action profile corresponds to a pair (a, S) , where a is the outcome obtained by the action profile and S is the set of types that outside observers believe are possible. An equilibrium is then a strategy profile in which the pairs (a, S) obtained in equilibrium are preferred

⁴Section 6.3 discusses this extension.

by all agents to pairs (b, T) obtained by unilateral deviations. But where do the sets of possible types come from? Suppose that agents play a strategy profile s , and some action profile is realized. Then the set of possible types consists of all type profiles R that could have led to the realized action profile.

In the majority example, the profile in which agents always vote truthfully leads to full revelation of information, since any realized action profile uniquely identifies all agents' types. However, this is not an equilibrium. To see this, suppose all agents are of type t_0 . Then one agent's unilateral deviation will not change the outcome, since the majority will still be 0. However, a unilateral deviation by an agent will lead to a different set of possible types in which outside observers incorrectly believe this agent is of type t_1 . Agents prefer deception over revelation, and so this deviation is profitable.

An additional problem with the simple voting mechanism is that it is not a full implementation—there are other strategy profiles that are equilibria but that do not yield the majority. A simple example is the profile in which all agents always vote for 0.

To address both of these difficulties, one might turn to more-complex mechanisms. Maskin (1999) shows that under some conditions that are satisfied in the majority example, his mechanism is a full implementation. But is it a full implementation also with information-sensitive preferences? It turns out that truthfulness is an equilibrium of this mechanism. However, whether or not there are other, undesirable equilibria depends on some aspects of agents' information-sensitive preferences that I have not yet specified. While agents prefer privacy over revelation when the outcome is *fixed*, what is their preferences when it is not? In particular, does type t_i prefer outcome i with revelation of information or outcome $1 - i$ with privacy? If he prefers the latter, then full implementation is impossible, and there will always be equilibria that do not yield the majority. This is formalized in Proposition 3.1, which shows that without restrictions on information-sensitive preferences, implementation by *any* mechanism may be impossible. If type t_i prefers outcome i with revelation, however, then the mechanism of Maskin (1999) is a full implementation. This is a consequence of Proposition 5.1. In fact, that proposition shows that if agents' preferences are such that they are willing to reveal their private information for some better outcome, then

full implementation is possible with information-sensitive preferences *whenever* it is possible with standard preferences.

Even if the mechanism of Maskin (1999) works, however, there is still a problem: In that mechanism, agents must reveal all their private information. But do there exist mechanisms for full implementation of majority that do not reveal information beyond the outcome? Theorem 4.1 shows that this is impossible.

However, not all is lost. If we allow for extensive-form mechanisms, in which communication proceeds in stages, then such privacy-protecting implementation of majority becomes possible. The possibility result is quite strong: Theorem 5.2 shows that with extensive-form mechanisms, privacy-protecting implementation is possible *whenever* implementation with standard preferences is possible.

Related Literature While in standard economic and game theoretic models agents have preferences only over material outcomes, there are two strands of the literature that consider agents who care also about the private information that is revealed. The literature on social image, including Bernheim (1994), Glazer and Konrad (1996), and Ireland (1994), studies the behavioral effects of agents' concerns for how they are perceived by others. The more general literature on psychological games studies agents who, in addition to being concerned with physical outcomes, also care about their beliefs and the beliefs of others (Geanakoplos et al. (1989)). However, while the modeling of agents in these literatures is related to this paper's model of agents with information-sensitive preferences, their aims are very different. In particular, these areas of research are typically not concerned with the problem of designing mechanisms for such agents, but rather with the study of various behavioral phenomena resulting from such agents' preferences.

Preferences that are not only over outcomes do appear in two implementation frameworks. Glazer and Rubinstein (1998) study an implementation problem in which agents may be motivated not only by the material outcomes of their actions, but also by the desire to have their own recommendation accepted. Matsushima (2008a,b) and Dutta and Sen (2012) study the problem of full implementation when agents have a strict preference for *honesty* when the resulting material outcome is not worsened.

The desire to control information transmission is also present endogenously and implicitly in many models of repeated interactions. In such interactions, agents play

a strategy that balances between myopic payoff maximization and information revelation that may potentially lead to a lower payoff in the future. This theme is prevalent in the literatures on the economics of privacy Acquisti et al. (2016) and on dynamic mechanism design (c.f. Vohra, 2012). The current paper is related in that it is also partly motivated by agents' concerns about revealing information that may impact later interactions. However, it is different in that it models such concerns exogenously, taking the position that it is often both impractical and unrealistic to model all future interactions. Furthermore, almost none of the work in those respective literatures considers a design perspective, but rather analyses fixed interactions.

Finally, this paper is related to a vast literature on privacy in computer science, particularly to privacy concerns in cryptography and to the newer study of differential privacy (Dinur and Nissim (2003)). These have been applied to strategic settings: for example, Naor et al. (1999) design a cryptographic system for guaranteeing privacy in auctions, and McSherry and Talwar (2007) utilize the tools of differential privacy to design economic mechanisms. However, in these applications, agents are not modeled as caring about privacy. That is, all these applications achieve some goal privately, but only when agents do not care about privacy. Once agents have preferences that depend on the information that is revealed, these applications can break down.

A number of more recent works do model agents' predilection for privacy explicitly by including a cost incurred by agents when some or all of their information is revealed. Miltersen et al. (2009) focus on the cryptographic implementation of a first-price auction when agents have marginal privacy concerns. Ghosh and Roth (2011) study the design of markets for selling privacy. Xiao (2011) studies the question of whether differential privacy is sufficient for truthful revelation of private information. Nissim et al. (2012) and Chen et al. (2011) consider a more general mechanism design problem in which agents are concerned about the information leaked by the outcome of a mechanism, and where all communication is hidden by perfect cryptography or a trusted third party.

While the motivation for these papers is similar to some of the motivation underlying the current paper, they are quite distinct. These papers all assume that communication is perfectly hidden by cryptography, and there are many situations in which the use of such technology is not feasible. For example, the US Senate and

House of Representatives conduct many recorded votes each year, in which the votes of all participants are publicized. When a buyer makes an online purchase, the vendor is generally aware of this purchase. When an individual testifies in a court of law, his testimony is heard by all present. In such cases many of the cryptographic tools are not applicable.

Furthermore, Nissim et al. (2012) and Chen et al. (2011) utilize tools from differential privacy to obtain mechanisms that do not reveal much of the agents' private information. There are two inherent features of such tools that make them inapplicable in many settings. First, they only apply when the number of agents is very large. Second, these mechanisms are randomized and do not always correctly implement the social choice function. In this sense they are close to the notion of virtual implementation of Abreu and Sen (1991). Unlike that notion, however, the approximation obtained is not arbitrarily close, but rather becomes close only as the number of agents becomes large. For a fixed number of agents, the mechanisms of Nissim et al. (2012) and Chen et al. (2011) could yield a suboptimal outcome with non-negligible probability.

Finally, in terms of the setting for implementation, the works of Nissim et al. (2012) and Chen et al. (2011) are incomparable to the current paper. The former consider partial implementation in dominant strategies in a setting with incomplete information and cardinal preferences, whereas the current paper examines full implementation in Nash and subgame perfect equilibrium in a setting with complete information and ordinal preferences.

Organization The rest of the paper is organized as follows. Section 2 presents the model. Sections 3 and 4 contain impossibility results—in the former I show that without restrictions on agents' privacy concerns there can be no implementation, and in the latter I show that, even with restrictions, almost nothing can be implemented in a privacy-protecting manner with normal-form mechanisms. Section 5 contains possibility results on information-sensitive and privacy-protecting implementation, and Section 6 contains further extensions of the model and results. Finally, the Appendix contains all proofs that do not appear in the main body of the text.

2 The Model

N denotes a finite set of agents and also, with a slight abuse of notation, its cardinality. \mathcal{O} denotes a possibly infinite set of outcomes.

2.1 Preferences and Social Choice Correspondences

We begin with the usual setup of agent preferences, but because these standard preferences will be extended we refer to them as the *intrinsic* preferences. The intrinsic preferences of an agent i are represented by a complete, transitive, binary relation R_i over \mathcal{O} , where aR_ib if agent i weakly prefers outcome $a \in \mathcal{O}$ over outcome $b \in \mathcal{O}$. Strict preference is denoted by P_i , and $\text{TR}_i(R) = \{a \in \mathcal{O} : aR_ib \ \forall b \in \mathcal{O}\}$ is the set of top-ranked alternatives for i under R . Denote by $R = (R_1, \dots, R_N)$ an intrinsic preference profile, and by (R_{-i}, \bar{R}_i) the intrinsic preference profile R in which agent i 's intrinsic preferences are replaced by \bar{R}_i . Finally, denote by \mathcal{R} the set of admissible profiles of intrinsic preferences.

A social choice correspondence (SCC) $F : \mathcal{R} \rightarrow \mathcal{O}$ is a mapping from a profile R of intrinsic preferences to a set of outcomes. An SCC is called a social choice function (SCF) if the range is always a singleton, i.e., if $F : \mathcal{R} \mapsto \mathcal{O}$. In this case denote the function by a lower case f . Denote by $F(\mathcal{R})$ the range of F when the domain is \mathcal{R} , namely $F(\mathcal{R}) = \cup_{R \in \mathcal{R}} F(R)$. Finally, an SCC F is *constant* if there exists some $a \in \mathcal{O}$ such that $a \in F(R)$ for all $R \in \mathcal{R}$, and otherwise F is *non-constant*.

For a given SCC F and an outcome $a \in \mathcal{O}$, denote by $\mathcal{R}|_{a,F} = \{R \in \mathcal{R} : a \in F(R)\}$. This is the set of profile preferences for which a is a possible outcome under F . When F is clear from context I will denote this set as $\mathcal{R}|_a$.

In this paper I extend preferences by adding privacy concerns for the agents. Agents' information-sensitive preferences depend not only on R , but also on a privacy state ψ . More formally, each agent i has preferences R_i^ψ that extend his intrinsic preferences R_i , where R_i^ψ is a complete, transitive, binary relation over $\mathcal{O} \times 2^{\mathcal{R}}$. The first coordinate is an element of \mathcal{O} , the outcome, and the second coordinate is a subset of \mathcal{R} , which I call the *set of possible types*. The set of possible types is the set of *intrinsic* preference profiles that outside observers believe are possible. For example, if a run of a mechanism reveals no information about the intrinsic preferences, then

the set of possible types is all of \mathcal{R} . If a run of a mechanism only reveals that the true intrinsic preferences R are such that a social choice function f satisfies $f(R) = a$, then the set of possible types is $\mathcal{R}|_a$. If a run of a mechanism reveals the true intrinsic preferences to be some $R \in \mathcal{R}$, then the set of possible types is $\{R\}$. The set of possible types thus represents the possibilistic beliefs (see, e.g. Aumann, 1976; Chen and Micali, 2015) of outside observers about the types of agents.

Now, for any agent i , any $S, T \subseteq \mathcal{R}$, and any $a, b \in \mathcal{O}$, it holds that $(a, S) R_i^\psi (b, T)$ if and only if agent i weakly prefers outcome a and set of possible types S over outcome b and set of possible types T . Denote by P_i^ψ the strict part of R_i^ψ .

In order to relate the information-sensitive preferences of an agent to his intrinsic preferences, I will assume that the former are equal to the latter when the set of possible types—and so the information disclosed—is fixed. That is, for every set of possible types S and $a, b \in \mathcal{O}$ it holds that $(a, S) R_i^\psi (b, S)$ if and only if aR_ib .⁵ Importantly, however, it may not be the case that $(a, S) R_i^\psi (b, T)$ for $T \neq S$, and of course the outcome will partly determine the set of possible types that is realized.

Now, there may be more than one possible privacy state, so denote by Ψ the set of all admissible privacy states. Then in our model, the following will hold: For any $i \in N$, any $\psi, \psi' \in \Psi$, any $R \in \mathcal{R}$, and any $S \subseteq 2^{\mathcal{R}}$ it will be the case that $(a, S) R_i^\psi (b, S)$ if and only if $(a, S) R_i^{\psi'} (b, S)$, which holds whenever aR_ib . However, for $S \neq T$ it may be the case that $(a, S) P_i^\psi (b, T)$ but $(b, T) P_i^{\psi'} (a, S)$. Note that in some of the discussion and results we will restrict our attention to the case where Ψ is a singleton, for simplicity.

Observe that this extended framework is a strict generalization of the standard framework, since one can model agents as not caring about privacy. Any agent i can express an unconditional preference of outcome a over b by preferring the pair (a, S) over (b, T) for all sets S and T . Denote by o the privacy state in which this is the case. Formally, for every $i \in N$, every $R \in \mathcal{R}$, every $S, T \subseteq \mathcal{R}$, and any $a, b \in \mathcal{O}$, it holds that $(a, S) R_i^o (b, T)$ if and only if aR_ib .

In the remainder of the paper I will sometimes refer to R^ψ as the state.

⁵For most of the results in this paper weaker assumptions suffice. One example of such an assumption is to associate intrinsic preferences with information-sensitive preferences when the set of possible types is $S = \{R\}$.

2.2 Mechanisms

A mechanism is an extensive-form game, together with a mapping from terminal histories to elements of \mathcal{O} . Formally:

Definition 2.1 (mechanism) *An N -person mechanism is a tuple (H, A, g) where*

- *H is a set of (finite) history sequences such that the empty word $\epsilon \in H$. A history $h \in H$ is terminal if $\{a : (h, a) \in H\} = \emptyset$. The set of terminal histories is denoted Z .*
- *$A = (A_1, \dots, A_N)$, where each A_i is a function that, for every non-terminal history $h \in H \setminus Z$, assigns a set $A_i(h)$ of actions available to agent i , where $(h, a) \in H$ for all $a \in A_1(h) \times \dots \times A_N(h)$.*
- *$g : Z \mapsto \mathcal{O}$ is a function that maps terminal histories to outcomes.*

A mechanism is called a normal-form mechanism if all non-empty histories are terminal.

2.3 Strategies

A strategy s_i of agent i in a mechanism (H, A, g) is a function that maps any pair (R^ψ, h) to an action from $A_i(h)$. Denote by $s = (s_1, \dots, s_n)$ a profile of strategies, by $s(R^\psi)$ the profile of strategies in state R^ψ , and by $H(s(R^\psi))$ the terminal history reached when the profile s is played in state R^ψ .

An R^ψ -deviation from a strategy s_i by an agent i is a strategy s'_i that agrees with s_i in every state except R^ψ . Formally, $s'_i(\bar{R}^\psi, h) = s_i(\bar{R}^\psi, h)$ for all h and $\bar{R}^\psi \neq R^\psi$. In state R^ψ the strategy s'_i may be different from s_i . Denote by (s'_i, s_{-i}) the strategy profile in which agent i plays s'_i and agents $j \neq i$ play s_j . Also, denote by $s(R^\psi)|_h$ the profile of strategies s in the subgame rooted at h when the preference profile is R^ψ , and by $H(s(R^\psi)|_h)$ the terminal history reached with this profile in the subgame rooted at h .

2.4 Equilibria

We begin with an informal description of equilibria in normal-form mechanisms. This is followed by formal definitions for both normal-form and extensive-form notions.

In the usual setup, each pure strategy profile corresponds to an outcome dictated by the mechanism. A (pure) Nash equilibrium is then a profile in which no agent has a profitable unilateral deviation. In our setup, however, each strategy profile will correspond to a pair—an outcome a and a set of possible types S . An equilibrium will then be a strategy profile in which no agent can unilaterally deviate to obtain a more favorable pair (b, T) .

The outcomes a and b are determined by the function g of the mechanism. But where do the sets of possible types S and T come from? The following is motivated by the notion of a perfect Bayesian equilibrium. Suppose that in some mechanism agents play a strategy profile s , that $\Psi = \{\psi\}$ (that is, it is known to outside observers that the privacy state is ψ), and that some action profile is realized. Then the set of possible types at this action profile is

$$\{R : s(R^\psi) \text{ leads to the realized action profile}\}.$$

That is, on the path of a profile s , the set of possible types is precisely the set of preferences that lead to the realized action profile.

Next, what is the set of possible types at action profiles that cannot be reached by the strategy profile s ? This question is similar to the question of off-equilibrium beliefs in perfect Bayesian equilibrium. As in that equilibrium concept, in most of this paper the sets of possible types at these action profiles will be part of the description of a particular equilibrium. In Section 6.1 I will argue that the possibility results of this paper are robust to a relaxation of this notion.

To formalize the discussion above fix a normal-form mechanism (H, A, g) and suppose that the agents play a strategy profile s and that the terminal node reached is some $z \in Z$. Then define the set

$$L(z, s) = \{R \in \mathcal{R} : H(s(R^\psi)) = z \text{ for some } \psi \in \Psi\}.$$

Next, define the belief $\beta : \mathcal{O} \mapsto 2^{\mathcal{R}}$, a function from outcomes to subsets of \mathcal{R} , the set of possible types. β is *s-consistent* if for each z , $\beta(z) = L(z, s)$ whenever $L(z, s) \neq \emptyset$.

We now define one notion of equilibrium that we will use: A variant of Nash equilibrium, broadened to allow for preferences over both outcomes and sets of possible types.

Definition 2.2 (information-sensitive Nash equilibrium) *A strategy-belief pair (s, β) in a mechanism (H, A, g) is an information-sensitive Nash equilibrium at Ψ if for every $i \in N$, $R \in \mathcal{R}$, $\psi \in \Psi$, and R^ψ -deviation s'_i of agent i ,*

$$(g(z), \beta(z)) \ R_i^\psi \ (g(z'), \beta(z')),$$

where $z = H(s(R^\psi))$ and $z' = H((s'_i, s_{-i})(R^\psi))$.

We will often require that in an equilibrium (s, β) the beliefs β are s -consistent (analogously to perfect Bayesian equilibria), but will not require this as part of the definition to facilitate our notion of implementation. In particular, this is relevant for bullet 2 of Definition 2.6.

We now define the subgame perfect variant of Definition 2.2. For this, suppose the mechanism (H, A, g) is not a normal-form mechanism. We first extend the notion of beliefs β to a dynamic setting by defining sets β_h for every nonterminal $h \in H$.

Suppose that the agents play a strategy profile s and that the terminal node reached is some $z \in Z$. Then define the sets

$$L_h(z, s) = \{R \in \mathcal{R} : H(s(R^\psi)|_h) = z \text{ for some } \psi \in \Psi\}.$$

Next, define the belief β as a function from outcomes, strategies, and histories to subsets of \mathcal{R} . As before, β is s -consistent if for each z , s , and h , the set $\beta_h(z)$ is a subset of \mathcal{R} that satisfies $\beta_h(z) = L_h(z, s)$ whenever $L_h(z, s) \neq \emptyset$.

We now define the second notion of equilibrium that we will use. Observe that this definition, when applied to a normal-form mechanism, is equivalent to Definition 2.2.

Definition 2.3 (information-sensitive subgame perfect equilibrium) *A strategy-belief pair (s, β) in a mechanism (H, A, g) is an information-sensitive subgame perfect equilibrium at Ψ if for every $i \in N$, $R \in \mathcal{R}$, $\psi \in \Psi$, nonterminal $h \in H$, and R^ψ -deviation s'_i of agent i ,*

$$(g(z), \beta_h(z)) \ R_i^\psi \ (g(z'), \beta_h(z')),$$

where $z = H(s(R^\psi)|_h)$ and $z' = H((s'_i, s_{-i})(R^\psi)|_h)$.

The relation between Definition 2.3 and the standard notion of a subgame perfect equilibrium (SPE) is the following: A pair (s, β) is an information-sensitive subgame perfect equilibrium at privacy state o if and only if for every profile of preferences $R \in \mathcal{R}$, the strategy profile $s(R^o)$ is a (standard) SPE.

2.5 Implementation

The standard setting A mechanism (H, A, g) is a subgame perfect implementation of an SCC F if for every $R \in \mathcal{R}$ the set of outcomes obtained by subgame perfect equilibria of (H, A, g) is equivalent to $F(R)$.

Abreu and Sen (1990) show that the following condition is necessary for implementation in SPE:

Definition 2.4 (Condition α) *An SCC F satisfies Condition α if for all $R, \bar{R} \in \mathcal{R}$ and outcomes $a \in F(R) - F(\bar{R})$ there exist a sequence of agents $j(0), \dots, j(\ell)$ and a sequence of outcomes $a = a_0, a_1, \dots, a_\ell, a_{\ell+1}$ such that*

- (i) $a_k R_{j(k)} a_{k+1}; k = 0, \dots, \ell$
- (ii) $a_{\ell+1} \bar{P}_{j(\ell)} a_\ell$
- (iii) $a_k \notin \text{TR}_{j(k)}(\bar{R})$ for $k = 0, \dots, \ell$
- (iv) if $a_{\ell+1} \in \text{TR}_i(\bar{R})$ for all $i \neq j(\ell)$, then either $\ell = 0$ or $j(\ell - 1) \neq j(\ell)$.

Abreu and Sen (1990) also show that Condition α , together with no veto power, is sufficient for subgame perfect implementation when there are 3 or more agents⁶.

Definition 2.5 (no veto power (NVP)) *An SCC F satisfies no veto power (NVP) if the following holds for every $R \in \mathcal{R}$ and $a \in \mathcal{O}$: if $a \in \text{TR}_i(R)$ for at least $N - 1$ agents i , then $a \in F(R)$.*

Implementation with privacy With privacy concerns, our definition of information-sensitive implementation of an SCF⁷ f (Definition 2.6 below) states that for every R^ψ , the outcome obtained by information-sensitive subgame perfect equilibria should always be $f(R)$. In order to facilitate the modification to privacy-protecting implementation, however, we split the definition into parts. Observe that Definition 2.6 is identical to subgame perfect implementation in the standard setting when the privacy state $\psi = o$.

⁶Vartiainen (2007) gives conditions for subgame perfect implementation that are both necessary and sufficient.

⁷I focus on implementations of SCFs and not SCCs for simplicity. In Section 6.2 I extend this discussion to SCCs.

Definition 2.6 (information-sensitive implementation) *A mechanism (H, A, g) is an information-sensitive subgame perfect implementation of an SCF f at Ψ if:*

1. *There exists a pair (s^*, β^*) for which the following hold:*
 - (a) *(s^*, β^*) is an information-sensitive subgame perfect equilibrium at Ψ and β^* is s^* -consistent, and*
 - (b) *$g(H(s^*(R^\psi))) = f(R)$ for all $R \in \mathcal{R}$ and $\psi \in \Psi$.*
2. *For all $R \in \mathcal{R}$, $\psi \in \Psi$, and pairs (s, β^*) that form an information-sensitive subgame perfect equilibrium at Ψ , it holds that $g(H(s(R^\psi))) = f(R)$.*

Observe that the equilibria in bullet 2 of Definition 2.6 are with respect to β^* , the set of possible types that is s^* -consistent. The observer knows which strategy profile is supposed to be played—namely, s^* —and has “beliefs” about the set of possible types corresponding to each history of play. Even if the agents play a different strategy profile such as s , this is unobserved by the observer, and thus he does not alter his sets of possible types. In Section 6.1 I provide a stronger definition of implementation that dispenses with this assumption, and for which the possibility results still hold.

Next, since we wish to design mechanisms that also protect agents’ privacy, we add the following element to our implementations.

Definition 2.7 (privacy-protecting implementation) *A mechanism (H, A, g) is a privacy-protecting implementation of an SCF f if it is an information-sensitive implementation of f , and if the strategy profile s^* guaranteed in Definition 2.6 also satisfies the following:*

1. (c) *For all $\psi \in \Psi$ and $R, \bar{R} \in \mathcal{R}$ such that $f(R) = f(\bar{R})$, there exists $\phi \in \Psi$ such that $H(s^*(R^\psi)) = H(s^*(\bar{R}^\phi))$.*

If $\Psi = \{\psi\}$, then condition 1(c) states that whenever preference profiles R and \bar{R} lead to the same outcome, the histories reached by $s^*(R^\psi)$ and by $s^*(\bar{R}^\psi)$ must be identical. For more general Ψ , condition 1(c) states that the terminal history $H(s^*(R^\psi))$ could have been reached by s^* with any profile of preferences \bar{R}^ϕ for which $f(\bar{R}) = f(R)$. Since the outside observer does not know which privacy state

from Ψ is the actual one, when he sees the outcome $H(s^*(R^\psi))$ he cannot differentiate between the true intrinsic preferences being R or \bar{R} .

A stronger variant of this definition might require that $H(s^*(\mathcal{R}^\psi)) = H(s^*(\bar{\mathcal{R}}^\phi))$ for all $\psi \in \Psi$ and $R, \bar{R} \in \mathcal{R}$ such that $f(R) = f(\bar{R})$. The possibility results of Theorems 5.2 and 5.5, in fact, satisfy this stronger variant.

3 Restrictions on Preferences

This section discusses restrictions on information-sensitive preferences. To motivate the necessity of such restrictions I first show that without them there may be no information-sensitive implementation.

Proposition 3.1 *Fix any set \mathcal{R} of intrinsic preferences and $\Psi = \{\psi\}$. For each $i \in N$ and $R \in \mathcal{R}$, let R^ψ satisfy the following: For any $a, b \in \mathcal{O}$ and set $S \subsetneq \mathcal{R}$, it holds that $(a, \mathcal{R})P_i^\psi(b, S)$. Then at Ψ there does not exist an information-sensitive subgame perfect implementation of any non-constant SCF.*

The difficulty with these preferences is that regardless of the outcome, each agent strictly prefers *no* information to be revealed than *some* information to be revealed. Thus, there will always be an equilibrium in which agents act as if they prefer some arbitrary outcome, regardless of their true preferences.

Restrictions on information-sensitive preferences Because of the impossibility of implementation with information-sensitive preferences implied by Proposition 3.1, we will restrict the preferences of agents. We will consider two main restrictions. The first is a strong notion, lexicographic preferences: Roughly speaking, preferences are lexicographic if agents care about privacy only insofar as the outcomes are unaffected. In other words, agents are willing to forego all privacy if they can obtain a more favorable outcome. However, if the outcome is unaffected, they may prefer to reveal as little information as possible.⁸

Definition 3.2 (lexicographic preferences) Ψ is lexicographic if for any $i \in N$, $R \in \mathcal{R}$, $\psi \in \Psi$, and sets $S, T \subseteq \mathcal{R}$, it holds that $(a, S)P_i^\psi(b, T)$ whenever $aP_i b$.

⁸This is similar in spirit to Dutta and Sen (2012), in which agents have a lexicographic preference for honesty.

The second restriction, called minimal willingness to reveal (MWR), is weaker. Roughly speaking, preferences satisfy MWR if agents are willing to reveal all their information for *some* outcome, and particularly for any top-ranked outcome.

Definition 3.3 (MWR preferences) Ψ satisfies minimal willingness to reveal (MWR) if for any $i \in N$, $R \in \mathcal{R}$, $\psi \in \Psi$, and sets $S, T \subseteq \mathcal{R}$, it holds that $(a, S)P_i^\psi(b, T)$ whenever $a \in \text{TR}_i(R)$ and $b \notin \text{TR}_i(R)$.

For implementations that are privacy-protecting and not only information-sensitive, we will need one more restriction on the preferences of agents. This restriction essentially states that for any outcome that is implemented, agents weakly prefer full privacy over full revelation of information.

Definition 3.4 (privacy favoring) Ψ is privacy favoring with respect to an SCF f if for each $\psi \in \Psi$, $i \in N$ and $R \in \mathcal{R}$ it holds that

$$(a, \mathcal{R}|_{a,f}) R_i^\psi(a, \{R\}),$$

where $a = f(R)$.

By Proposition 3.1, in order to have some possibility results agents should be willing to reveal some information to attain a better outcome. The weakest assumption is that that they are willing to reveal information to attain their top-ranked outcome, as in MWR preferences. But note that MWR preferences are still somewhat restrictive, as they require agents to willingly reveal *all* their information to attain a top-ranked outcome.

4 Implementation with Normal-Form Mechanisms

Maskin (1999) shows that a particular monotonicity condition, now called Maskin monotonicity, is necessary for implementation with normal-form mechanisms. As this condition is known to be quite restrictive (c.f. Muller and Satterthwaite, 1977; Saijo, 1987), there are various relaxations of the problem that render the theory more widely applicable. In particular, the relaxations include restricting the domain of preferences, considering a binary outcome space, and examining SCCs rather than

SCFs (Postlewaite and Wettstein, 1989; Serrano, 2004). In this section I show that even without requiring Maskin monotonicity, and even if we allow the first two of these relaxations, then implementation is once again nearly impossible if we require some privacy. In Section 6.2 I extend the impossibility result also to the case of SCCs.

The impossibility result in this section requires some assumptions on the information-sensitive preferences—in particular, that there are not too many dependencies between the agents’ preferences. This does not exclude the possibility of privacy-protecting implementation with normal-form mechanisms that take advantage of specific dependencies.⁹ However, the main point of the result is to contrast the impossibility with the general possibility using extensive-form mechanisms (Section 5).

We begin with three definitions. First, a domain \mathcal{R} is *independent* if for every privacy state $\psi \in \Psi$, agent $i \in N$, and preference profiles $R, \bar{R} \in \mathcal{R}$, there exists $\hat{R} \in \mathcal{R}$ such that $\hat{R}_i^\psi = R_i^\psi$ and $\hat{R}_{-i}^\psi = \bar{R}_{-i}^\psi$. Second, a set of privacy states Ψ is *independent* if for all agents $i \in N$, preference profiles $R, \bar{R} \in \mathcal{R}$ such that $R_i = \bar{R}_i$, and privacy states $\psi \in \Psi$, it holds that $R_i^\psi = \bar{R}_i^\psi$. Third, a set of privacy states Ψ is *convex* if for every pair of privacy states $\psi, \psi' \in \Psi$, agent $i \in N$, and preference profile $R \in \mathcal{R}$, there exists $\hat{\psi} \in \Psi$ such that $R_i^{\hat{\psi}} = R_i^\psi$ and $R_{-i}^{\hat{\psi}} = R_{-i}^{\psi'}$. A particularly simple example of a convex Ψ is the singleton, $\Psi = \{\psi\}$.

The impossibility result follows. Note that, while it does assume independence, it does not assume that Ψ is lexicographic or even MWR.

Theorem 4.1 *Fix any independent domain \mathcal{R} for which there exists $a \in \mathcal{O}$ and $R \in \mathcal{R}$ such that $a \in \text{TR}_i(R)$ for all $i \in N$, as well as any non-constant SCF $f : \mathcal{R} \mapsto \mathcal{O}$ that satisfies NVP. Then there is no normal-form mechanism that is a privacy-protecting implementation of f at any independent, convex Ψ .*

Note that the conditions of the theorem are quite weak, as they require only the *existence* of some preference profile in which all agents agree on the top-ranked outcome. In particular, the conditions are satisfied in the simple majority example from the introduction. Finally, the conditions for impossibility can be further weakened, as follows. Suppose there *exists* some $\mathcal{O}' \subseteq \mathcal{O}$ such that f satisfies NVP with respect to \mathcal{O}' : that is, if all but one agent prefer outcome $a \in \mathcal{O}'$ over all other outcomes *in* \mathcal{O}' ,

⁹It is well known that the use of certain dependencies can facilitate the design of mechanisms that would otherwise not be possible (e.g. Cremer and McLean, 1988).

then the outcome is a . Furthermore, suppose there is some $R \in \mathcal{R}$ under which all agents prefer $a \in \mathcal{O}'$ over other outcomes *in* \mathcal{O}' . Then the conclusion of Theorem 4.1 will hold.¹⁰

The intuition underlying Theorem 4.1 is the following. Suppose for simplicity that $\Psi = \{\psi\}$, and observe that in this case, the equilibrium strategy s^* of a normal-form mechanism for privacy-protecting implementation can depend only on the outcome $f(R)$ (and not on the full state R^ψ). Since f satisfies NVP, this means that at any profile R in which all agents except i prefer outcome a the most, the profile $s^*(a)$ (the profile played when the outcome is a) is an equilibrium. That is, under such R agent i weakly prefers to follow $s^*(a)$ regardless of his own preferences. But this implies that all agents always playing $s^*(a)$ regardless of the state is also an equilibrium—by the same logic, no agent has an incentive to deviate.

5 Implementation with Extensive-Form Mechanisms

In this section I show that information-sensitive and privacy-protecting implementation in subgame perfect equilibrium are possible using extensive-form mechanisms. For information-sensitive implementation I have the following proposition:

Proposition 5.1 *If there is a subgame perfect implementation of an SCF $f : \mathcal{R} \mapsto \mathcal{O}$ that satisfies NVP and $N \geq 3$, then there is an information-sensitive implementation of f for any lexicographic Ψ .*

The idea behind this proposition is to show that the mechanism of Moore and Repullo (1988), with slight modification, works also for information-sensitive agents. The reason is that in this mechanism, all private information is revealed, and so beliefs on the equilibrium path consist of the true profile of intrinsic preferences. Furthermore, this holds also off the equilibrium path, as long as not many agents deviate. Thus, in equilibrium it is as if the set of possible types is fixed, and so only the intrinsic preferences matter.

I then extend this proposition and prove the following theorem on the possibility of privacy-protecting implementation:

¹⁰In particular, under this weakening the impossibility result may hold also in economic environments.

Theorem 5.2 *If there is a subgame perfect implementation of an SCF $f : \mathcal{R} \mapsto \mathcal{O}$ that satisfies NVP and $N \geq 3$, then there is a privacy-protecting implementation of f for any lexicographic, privacy favoring Ψ .*

The proof of Theorem 5.2 is constructive—it describes a mechanism that is a privacy-protecting implementation of f . While the mechanism is a bit intricate, the main idea of the construction is the following. In the first stage of the mechanism, agents attempt to coordinate on an outcome only, without revealing any additional private information. If there is no unanimous agreement, then the mechanism proceeds with a “contingency plan”—essentially, this is the information-sensitive implementation from Proposition 5.1, in which there is full information revelation. The revelation of information here acts as a sort of “threat” against mis-coordination, and in equilibrium this contingency plan is never invoked.

If agents coordinate on an incorrect outcome in the first stage, the hope is that there is some agent who benefits from deviating, leading to the “correct” outcome (since the contingency plan is the mechanism from Proposition 5.1) and full revelation of his private information. If this were always the case, then the construction of the mechanism would have been straightforward. The intricacies of the mechanism result from the necessity of handling cases in which this is not the case.

Proposition 5.1 and Theorem 5.2 provide possibility results for lexicographic Ψ . A necessary condition for these implementations is that there be an SPE implementation of f . As Abreu and Sen (1990) show, a necessary condition for SPE implementation is Condition α (see Definition 2.4). For implementation with MWR preferences, as opposed to just lexicographic preferences, we need the following stronger condition:

Definition 5.3 (Condition α^{\max}) *An SCC F satisfies Condition α^{\max} if it satisfies Condition α but with (ii)' replacing (ii):*

$$(ii)' \quad a_{\ell+1} \in \text{TR}_{j(\ell)}(\bar{R}).$$

Condition α differs from Condition α^{\max} in item (ii)—the former requires a preference reversal, whereas the latter requires a preference reversal where one outcome becomes top ranked.

The following proposition and theorem are the counterparts to Proposition 5.1 and Theorem 5.2 for MWR preferences:

Proposition 5.4 *If $N \geq 3$ and the SCF $f : \mathcal{R} \mapsto \mathcal{O}$ satisfies Condition α^{\max} and NVP, then there is an information-sensitive implementation of f for any MWR Ψ .*

Theorem 5.5 *If $N \geq 3$ and the SCF $f : \mathcal{R} \mapsto \mathcal{O}$ satisfies α^{\max} and NVP, then there is a privacy-protecting implementation of f for any MWR, privacy favoring Ψ .*

Remark 5.6 The mechanisms used for the theorems in this section utilize integer games, which are quite unrealistic in real-life mechanisms. Note, however, that the use of such games in our setting is inherited from the mechanisms of Maskin (1999) and Moore and Repullo (1988) which this paper builds on, and without them the results here would probably not be so general. The design of general mechanisms that do not use such games is an open question even in the standard (no-privacy) setting, and is a research agenda orthogonal to this paper. However, I believe that, as in the works of Maskin (1999) and Moore and Repullo (1988), the ideas used in our mechanisms will lead to more-realistic mechanisms in more-specific contexts.

Remark 5.7 Note that the possibility results for privacy-protecting implementation with extensive-form mechanisms relies on the ability of agents to condition their actions on the preference profile R , in addition to the outcome $f(R)$, unlike the case of normal-form mechanisms. If one requires privacy also off the equilibrium path (formally, that for each history, the actions of agents at that history do not reveal additional information about the preferences), then the impossibility result of Theorem 4.1 will be reestablished also for extensive-form mechanisms. The reason is that in that case, the strategy of each agent will depend only on the outcome $f(R)$, as in the case of normal-form mechanisms. Without this requirement, the strategy in extensive-form mechanisms may also depend on R , as long as $s_i(R, h)$ depends only on $f(R)$ and h whenever h is on the equilibrium path.

6 Extensions

6.1 Robustness

One of the implicit assumptions in the definitions of information-sensitive and privacy-protecting implementation is that outside observers know the strategy profile being

played. That is, the second requirement of these definitions is that for all profiles s that form an information-sensitive SPE, the outcome should be the same as dictated by f . However, the notion of an information-sensitive SPE relies on the sets β , which of course depend on the profile s being played. Thus, a natural question is, what if outside observers do not know the profile s ? Perhaps they have a probabilistic belief about the profile s being played, or perhaps their uncertainty over which s is played is Knightian. In such situations, we may want a stronger notion of implementation. In particular, we may want the following requirement: for any profile s being played, if the outcome is not the same as dictated by f , then some player should have a profitable deviation from this profile *regardless of the beliefs of outside observers about the profile being played*. This is formally captured by bullet 2 of the following definition.

Definition 6.1 (strong information-sensitive implementation) *A mechanism (H, A, g) is a strong privacy-protecting subgame perfect implementation of an SCF f at Ψ if:*

1. *There exists a pair (s^*, β^*) for which the following hold:*
 - (a) *(s^*, β^*) is an information-sensitive subgame perfect equilibrium at Ψ and β^* is s^* -consistent, and*
 - (b) *$g(H(s^*(R^\psi))) = f(R)$ for all $R \in \mathcal{R}$.*
 - (c) *For all $R, \bar{R} \in \mathcal{R}$ such that $f(R) = f(\bar{R})$, it holds that $H(s^*(\mathcal{R}^\psi)) = H(s^*(\bar{\mathcal{R}}^\psi))$.*
2. *For all $R \in \mathcal{R}$ and strategy profiles s for which $g(H(s(R))) \neq f(R)$, there exists an agent $i \in N$, a history $h \in H$, and an R^ψ -deviation s'_i of agent i such that*

$$(g(H((s'_i, s_{-i})(R^\psi)|_h)), T) P_i^\psi (g(H(s(R^\psi)|_h)), S)$$

for all sets $S, T \subseteq \mathcal{R}$.

Note that the proofs of Propositions 5.1 and 5.4, as well as Theorems 5.2 and 5.5 go through with this stronger definition of implementation.

6.2 Social Choice Correspondences

In this section I extend the notion of privacy-protecting implementation to SCCs. We first need a definition.

Definition 6.2 (restrictions of an SCC) *A restriction of an SCC $F : \mathcal{R} \rightarrow \mathcal{O}$ is any function from the set $\{f : \mathcal{R} \mapsto \mathcal{O} \text{ such that } f(R) \in F(R) \forall R \in \mathcal{R}\}$.*

Consider the following definition, which is the analog of Definition 2.7 for the case of SCCs. First, modify Definition 2.6 so that bullet 2 applies to SCCs rather than SCFs, as follows: For all $R \in \mathcal{R}$, $\psi \in \Psi$, and pairs (s, β^*) that form an information-sensitive subgame perfect equilibrium at Ψ , it holds that $g(H(s(R^\psi))) \in F(R)$. Next, modify Definition 2.7 accordingly, with bullet 2 modified for SCCs. Finally:

Definition 6.3 (privacy-protecting implementation of an SCC) *A mechanism (H, A, g) is a privacy-protecting subgame perfect implementation of an SCC F at Ψ if it is a privacy-protecting subgame perfect implementation of every restriction f of F .*

I first argue that the impossibility result of Theorem 4.1 holds also for SCCs. Fix any non-constant SCC F that satisfies NVP, and such that there exists some $R \in \mathcal{R}$ and $a \in \mathcal{O}$ satisfying $a \in \text{TR}_i(R)$ for all $i \in N$. Construct the following restriction f of F : For each $\bar{R} \in \mathcal{R}$ such that there exists $j \in N$ such that $a \in \text{TR}_i(\bar{R})$ for all $i \neq j$, fix $f(\bar{R}) = a$. For all other $\bar{R} \in \mathcal{R}$, fix $f(\bar{R}) = b$ for some arbitrary $b \in F(\bar{R})$. Then f is non-constant, and “almost” satisfies NVP: At profiles \bar{R} in which all but one agent (weakly) prefer outcome a , the outcome of f will be a . This property of “almost” NVP suffices for the proof of Theorem 4.1 to hold at f . Furthermore, the impossibility result goes through also for constant SCCs, as long as the outcome a commonly preferred by all agents under R does not satisfy $a \in f(R)$ for all $R \in \mathcal{R}$.

Second, the possibility results of Section 5 essentially go through for SCCs without modification.

6.3 Information-Limiting Implementation

For privacy-protecting implementation the goal is to design a mechanism in which the only information revealed is the outcome, and not any additional private information

beyond that. In this section we explore a more general notion in which agents may wish to reveal some or all information.

First, observe that it is impossible to reveal any *less* information than what is revealed in privacy-protecting implementation, while at the same time correctly implementing an SCF. This is simply because the correct outcome, together with the knowledge that it is the correct outcome, imply something about agents' preferences—namely, that the preferences are such that the SCF yields whatever outcome was implemented. Next, observe that the most information that can be revealed is the entire set of preferences R . Thus, in this section we will consider implementation with revelation of information that is anywhere from being as fine as full revelation to as coarse as only revealing the outcome.

Now, it could be that in some profile of preferences R , an agent wants revelation of information, whereas in another profile \bar{R} he desires privacy—i.e., the observer is unable to distinguish between R and \bar{R} . However, these are clearly impossible goals to achieve simultaneously (if in state R the observer learns the true state, then if he does not learn it he can deduce that the state is *not* R , violating the privacy desideratum). What can be achieved is the revelation and concealment of information according to a partition of \mathcal{R} . To that end, we will utilize the following notation: We will denote by Π a partition of \mathcal{R} and by $\Pi(R) \subseteq \mathcal{R}$ the element of Π that includes R .

Definition 6.4 (information-limiting implementation) *A mechanism (H, A, g) is an information-limiting implementation of an SCF f with respect to a partition Π if it is an information-sensitive implementation of f , and if the strategy profile s^* guaranteed in Definition 2.6 also satisfies the following:*

1. (c) *For all $R \in \mathcal{R}$ it holds that if $\bar{R} \in \Pi(R)$, then $H(s^*(\mathcal{R}^\psi)) = H(s^*(\bar{R}^\psi))$, but if $\bar{R} \notin \Pi(R)$, then $H(s^*(\mathcal{R}^\psi)) \neq H(s^*(\bar{R}^\psi))$.*

Condition 1(c) states that the terminal history $H(s^*(\mathcal{R}^\psi))$ could have been reached by s^* with any profile of preferences \bar{R}^ψ for which $\bar{R} \in \Pi(R)$. Thus, the planner or any outside observer who sees the outcome $H(s^*(\mathcal{R}^\psi))$ cannot differentiate between the true intrinsic preferences being R or \bar{R} . Unlike the case of privacy-protecting implementation, however, the planner can differentiate between R and \bar{R} satisfying $f(R) = f(\bar{R})$ if $\bar{R} \notin \Pi(R)$.

Before stating the theorem we need a couple of definitions. The first is an appropriate variant of privacy favoring ψ .

Definition 6.5 (Π -favoring) Ψ is Π -favoring if for each $\psi \in \Psi$ $i \in N$ and $R \in \mathcal{R}$, it holds that

$$(a, \Pi(R)) R_i^\psi (a, \{R\}).$$

The next definition requires that the partition be a refinement of the partition $\{\mathcal{R}|_{a,f}\}_{a \in \mathcal{O}}$. As discussed above, this is a necessary condition for implementation.

Definition 6.6 (f -consistent) A partition Π is f -consistent if for each $R \in \mathcal{R}$ and $\bar{R} \in \Pi(R)$, it holds that $f(R) = f(\bar{R})$.

The information-limiting variant of Theorem 5.2 is the following:

Theorem 6.7 *If there is a subgame perfect implementation of an SCF $f : \mathcal{R} \mapsto \mathcal{O}$ that satisfies NVP and $N \geq 3$, then there is an information-limiting implementation of f with respect to an f -consistent partition Π for any lexicographic, Π -favoring ψ .*

The information-limiting variant of Theorem 5.5 is the following:

Theorem 6.8 *If $N \geq 3$ and the SCF $f : \mathcal{R} \mapsto \mathcal{O}$ satisfies α^{\max} and NVP, then there is an information-limiting implementation of f with respect to an f -consistent partition Π for any MWR, Π -favoring Ψ .*

7 Conclusion

In this paper I have presented a simple model of agents with information-sensitive preferences, and have studied the problem of implementation for such agents. There are two interesting directions for future work. First, the SCCs I have studied depend only on the intrinsic preferences of agents. However, one might broaden the domain of such functions to include the full state – intrinsic preferences plus the privacy state. For example, one might consider SCCs that yield one outcome when agents do not care about privacy, and another when they do, even if the intrinsic preferences are the same in both cases. The second direction is, of course, to extend the current framework to one of incomplete information – see Gradwohl and Smorodinsky (2017) for a study games in such a model.

Appendix

A Proof of Proposition 3.1

Proof of Proposition 3.1: Fix an SCF $f : \mathcal{R} \mapsto \mathcal{O}$, and suppose towards a contradiction that the mechanism (H, A, g) is an information-sensitive subgame perfect implementation of some non-constant f at ψ . Since f is non-constant, there exist distinct $a, b \in \mathcal{O}$ and intrinsic preferences $R^a, R^b \in \mathcal{R}$ such that $f(R^a) = a$ and $f(R^b) = b$. Let s^* be the strategy profile guaranteed by Definition 2.6.

Consider now the strategy profile s , such that $s(R) \equiv s^*(R^a)$ for all $R \in \mathcal{R}$. Also, for every nonterminal history $h \in H$, let $\beta_h(H(s^*(R^a)|_h)) = \mathcal{R}$. In words, these sets of possible types mean that starting at any history, if a terminal history is reached that could have been reached by s , then all intrinsic preferences are possible. This is reasonable, since s is the same regardless of the intrinsic preferences. We will define $\beta_h(z)$ for other z 's in the sequel.

We claim that the profile s constitutes an information-sensitive subgame perfect equilibrium at ψ (when the sets β are as will be defined). This, of course, implies a contradiction, since s does *not* always yield the social optimum according to f in (H, A, g) . In particular, s always yields the outcome a , but there are intrinsic preferences in which the socially optimal outcome is b (namely, R^b).

We now show that s is an information-sensitive subgame perfect equilibrium at ψ . Suppose towards a contradiction that this is not the case, and that there is some nonterminal history h , some $i \in N$, some $R \in \mathcal{R}$, and some R^ψ -deviation s'_i of agent i for which

$$(g(z'), \beta_h(z')) P_i^\psi (g(z), \beta_h(z)),$$

where $z = H(s(R^\psi)|_h)$ and $z' = H((s'_i, s_{-i})(R^\psi)|_h)$.

Recall our assumption above that $\beta_h(z) = \mathcal{R}$, yielding

$$(g(z'), \beta_h(z')) P_i^\psi (g(z), \mathcal{R}).$$

The crucial question now is, what should $\beta_h(z')$ be?

Observe that regardless of $\beta_h(z')$, it must be the case that $g(z')P_i^\psi g(z)$. For

otherwise, if $g(z')R_i^\psi g(z)$, then it follows that

$$(g(z'), \beta_h(z')) R_i^\psi (g(z), \mathcal{R}),$$

since the only way to get a strict improvement over an outcome and no information revelation is to get a better outcome and no information revelation. However, now a planner can “learn” that agent i will deviate *only* if this yields him a strictly better outcome. In other words, the planner now learns that the true intrinsic preferences must lie in the set $\mathcal{R}' = \{R \in \mathcal{R} : g(z')P_i g(z)\} \subsetneq \mathcal{R}$! Thus, defining the set $\beta_h(z') \stackrel{\text{def}}{=} \mathcal{R}'$ yields the contradiction that

$$(g(z'), \mathcal{R}') P_i^\psi (g(z), \mathcal{R}).$$

Hence, s is an information-sensitive subgame perfect equilibrium. ■

B Proof of Theorem 4.1

Proof of Theorem 4.1: Suppose towards a contradiction that there is a privacy-protecting implementation of f at a convex Ψ , and let (H, A, g) be the implementing mechanism and s^* the strategy profile guaranteed by Definition 2.7.

Fix some $\psi \in \Psi$. Let $a \in \mathcal{O}$ and $\bar{R} \in \mathcal{R}$ be such that $a \in \text{TR}_i(\bar{R})$ for all $i \in N$, and observe that $g(H(s^*(\bar{R}^\psi))) = a$ by NVP. Now, since f is non-constant, there exists some profile $\hat{R} \in \mathcal{R}$ such that $f(\hat{R}) = b \neq a$. Consider an agent i and a profile ${}_iR$ that is identical to \bar{R} for all except agent i , namely ${}_iR_{-i} = \bar{R}_{-i}$, and such that ${}_iR_i = \hat{R}_i$. Such a profile ${}_iR$ is in \mathcal{R} since \mathcal{R} is independent. By NVP, at preference profile ${}_iR$ it also holds that $f({}_iR) = f(\bar{R}) = a$. Since (H, A, g) is privacy-protecting there is some $\psi_i \in \Psi$ for which $H(s^*({}_iR^{\psi_i})) = H(s^*(\bar{R}^\psi))$ (by condition 1(c) of Definition 2.7). Recall that (H, A, g) is a normal-form mechanism, and so this implies that $s^*({}_iR^{\psi_i})$ is identical to $s^*(\bar{R}^\psi)$. Since s^* is an equilibrium, for any ${}_iR^{\psi_i}$ -deviation s'_i by agent i it holds that

$$(a, \beta^*(z)) {}_iR_i^{\psi_i} (g(z'), \beta^*(z')), \tag{1}$$

where $z = H(s^*({}_iR^{\psi_i})) = H(s^*(\bar{R}^\psi))$ and $z' = H((s'_i({}_iR^{\psi_i}), s_{-i}^*))$.

Let ϕ be the privacy-state for which $\hat{R}_i^\phi = \hat{R}_i^{\psi_i}$ for all $i \in N$. Such a ϕ must belong to Ψ since the latter is convex. Consider the strategy profile s that is identical to s^* everywhere except at \hat{R}^ϕ , and set $s(\hat{R}^\phi) = s^*(\bar{R}^\psi)$.

I claim that s is an information-sensitive equilibrium with respect to β^* . It is clearly an equilibrium at every state except \hat{R}^ϕ , since s^* is an equilibrium. What about at \hat{R}^ϕ ? First note that $s(\hat{R}^\phi)$ leads to outcome a and set of possible types $\beta^*(z, s^*)$, where $z = H(s^*(\bar{R}^\psi)) = H(s^*({}_iR^{\psi_i}))$ for every i . Consider the guarantee provided by (1), and note that ϕ was chosen so that $\hat{R}_i^\phi = \hat{R}_i^{\psi_i}$. Any \hat{R}_i^ϕ -deviation \hat{s}_i by agent i leads to an outcome $z' = H((\hat{s}_i(\hat{R}^\phi), s_{-i})) = H((\bar{s}_i(\bar{R}^\psi), s_{-i}^*)) = H((s'_i({}_iR^{\psi_i}), s_{-i}^*))$, where \bar{s}_i and s'_i are the \bar{R}^ψ -deviation and ${}_iR^{\psi_i}$ -deviation of agent i , respectively, that satisfy $\bar{s}_i(\bar{R}^\psi) = s'_i({}_iR^{\psi_i}) = \hat{s}_i(\hat{R}^\phi)$.

Finally, recall that (1) guarantees that

$$(a, \beta^*(z)) \quad {}_iR_i^{\psi_i} \quad (g(z'), \beta^*(z')).$$

But ${}_iR_i^{\psi_i} = \hat{R}_i^{\psi_i}$ by the independence of Ψ , and $\hat{R}_i^{\psi_i} = \hat{R}_i^\phi$ by the construction of ϕ , and so

$$(a, \beta^*(z)) \quad \hat{R}_i^\phi \quad (g(z'), \beta^*(z')).$$

Thus, agent i has no incentive to deviate. Furthermore, the analysis above holds for any agent $i \in N$ and corresponding preference profile ${}_iR$. Thus, s is an equilibrium with respect to β^* . However, this is a contradiction to bullet 2 of Definition 2.6, since $g(H(s(\hat{R}^\phi))) = a \neq b = f(\hat{R})$. ■

C Proofs from Section 5

Our possibility results about implementation with extensive-form mechanisms are constructive, and the mechanism we use are variants of the mechanism of Moore and Repullo (1988).

The Moore-Repullo mechanism that implements an SCC F in subgame perfect equilibrium, and which is also used by Abreu and Sen (1990), uses sequences of agents $j(0), \dots, j(\ell)$ and a sequences of outcomes $a_0, \dots, a_{\ell+1}$, one such pair of sequences for each $R \in \mathcal{R}$, $\bar{R} \in \mathcal{R}$, and $a \in F(R) - F(\bar{R})$, satisfying Condition α (see Definition 2.4). The mechanism is the following (copied almost verbatim from Abreu and Sen (1990)):

The Moore-Repullo Mechanism (MR):

- **Stage 0:** Each agent i simultaneously submits a triplet $(R^i, a^i, n^i) \in \mathcal{R} \times \mathcal{O} \times \mathbb{Z}$. If $N - 1$ agents submit the same R and $a \in F(R)$ then the outcome is a , unless the non-agreeing agent j announces R^j with $a \in F(R) - F(R^j)$ and $j = j(0)$ in the sequence $j(R, R^j, a)$. In this latter case, go to Stage 1. In all other cases the agent who announced the highest integer selects any outcome in \mathcal{O} .

- **Stage k , $k = 1, \dots, \ell$:** Each agent i simultaneously either raises a “flag” or announces a nonnegative integer.

If at least $N - 1$ agents raise flags, the agent $j(k - 1)$ (in the sequence $j(R, R^j, a)$) selects any outcome in \mathcal{O} .

If at least $N - 1$ agents announce 0 the outcome is a_k , unless $j(k)$ does not announce 0, in which case go to the next Stage, or, if $k = \ell$, implement $a_{\ell+1}$.

In all other cases the agent who announced the highest integer selects any outcome in \mathcal{O} .

Abreu and Sen (1990) show that the following strategy profile s^{MR} is a SPE of MR: In Stage 0, $s^{MR}(R, \epsilon) = (R, a, 0)$, where $a \in F(R)$. In all subsequent Stages all agents always announce 0. They also prove the following as part of the proof of their Theorem 2.

Lemma C.1 *Suppose the true strategy profile is R , but all agents submit $R^i = \bar{R} \neq R$ and outcome $a \notin F(R)$ in Stage 0 of the Moore-Repullo mechanism. Then if agent $j(0)$ of the sequence $j(\bar{R}, R, a)$ deviates and submits $R^i = R$, then all SPE outcomes following this deviation are in $TR_{j(0)}(R)$.*

Consider now the following mechanism MR' , which is a slight variation of the Moore-Repullo mechanism:

Mechanism MR':

- **Stage 0:** Same as the Moore-Repullo mechanism.
- **Stage k , $k = 1, \dots, \ell$:** Same as the Moore-Repullo mechanism, except that in each Stage k , each agent i also submits a vector of preferences $R^i \in \mathcal{R}$.

The MR' mechanism is almost identical to the Moore-Repullo mechanism, except for the expanded action space in Stages 1 through ℓ . Note that these additions do not impact the outcome of the mechanism. Consider the strategy profile $s^{\text{MR}'}$ that is identical to s^{MR} , except that in all Stages following 0 the agents submit the true profile of preferences R (in addition to announcing 0). Observe that $s^{\text{MR}'}$ is an SPE of the mechanism MR'. We use mechanism MR', together with the strategy profile $s^{\text{MR}'}$, in the proof of Proposition 5.1. But first, we need some definitions and lemmas.

C.1 Some Technical Definitions and Lemmas

Recall that

$$L(z, s) = \{R \in \mathcal{R} : H(s(R^\psi)) = z \text{ for some } \psi \in \Psi\},$$

and that β is s -consistent if $\beta(z) = L(z, s)$ whenever $L(z, s) \neq \emptyset$. Now define the sets

$$\text{LD}(z, s) = \{R \in \mathcal{R} : H((s'_i, s_{-i})(R^\psi)) = z \text{ for some } \psi \in \Psi, i \in N, \text{ and strategy } s'_i \text{ of agent } i\}.$$

β is s +deviation-consistent if it is s -consistent, and if for each z it holds that $\beta(z) = \text{LD}(z, s)$ whenever $L(z, s) = \emptyset$ and $\text{LD}(z, s) \neq \emptyset$. In words, if a given outcome z cannot be reached by the strategy profile s but *can* be reached by s with a unilateral deviation by some player, then the belief at that outcome is the set of all intrinsic profiles that would lead to z with s and some unilateral deviation. Similarly, define

$$\text{LD}_h(z, s) = \{R \in \mathcal{R} : H((s'_i, s_{-i})(R^\psi)|_h) = z \text{ for some } \psi \in \Psi, i \in N, \text{ and strategy } s'_i \text{ of agent } i\}.$$

β is s +deviation-consistent if it is s -consistent and if, for each z , s , and h , the set $\beta_h(z) = \text{LD}_h(z, s)$ whenever $L_h(z, s) = \emptyset$ and $\text{LD}_h(z, s) \neq \emptyset$.

Note that beliefs β that are s +deviation-consistent are restricted on the equilibrium path and close to it (that is, one deviation away), but not elsewhere.

Definition C.2 (maximally-dispersed) A strategy profile s in a mechanism (H, A, g) is maximally-dispersed at Ψ if for every $\psi \in \Psi$, $i \in N$, $R \in \mathcal{R}$, $h \in H \setminus Z$, and R^ψ -deviation $s'_i \neq s_i$ of agent i it holds that

- (i) $\text{LD}_h(H((s'_i, s_{-i})(R^\psi)|_h), s) = \{R\}$ and
- (ii) $\text{LD}_h(H(s(R^\psi)|_h), s) = \{R\}$.

Definition C.3 (outcome-condensed) A strategy profile s in a mechanism (H, A, g) is outcome-condensed at Ψ if for every $\psi \in \Psi$, $i \in N$, $R \in \mathcal{R}$, $h \in H \setminus Z$, and R^ψ -deviation $s'_i \neq s_i$ of agent i it holds that

- (i) $\text{LD}_h(H((s'_i, s_{-i})(R^\psi)|_h), s) = \{R\}$ and
 - (ii)' $g(s(R^\psi)|_h) = g((s'_i, s_{-i})(R^\psi)|_h) = a$ for some $a \in \mathcal{O}$ and
- $$\text{L}_h(H(s(R^\psi)|_h), s) = \mathcal{R}|_a.$$

Lemma C.4 Let s° be a subgame perfect equilibrium of a mechanism (H, A, g) , and let s be the strategy profile satisfying $s_i(R^\psi, h) = s_i^\circ(R, h)$ for every $\psi \in \Psi$ and all i , R , and h . Then if either s is outcome-condensed at Ψ and Ψ is privacy favoring, or s is maximally-dispersed at Ψ , then (s, β) is also an information-sensitive subgame perfect equilibrium at Ψ for any s -deviation-consistent β .

Proof: Fix an agent i , a profile of intrinsic preferences $R \in \mathcal{R}$, a privacy state $\psi \in \Psi$, and some $h \in H \setminus Z$. We will show that agent i does not have a unilateral R^ψ -deviation from s at h that will yield him a strictly higher payoff.

Let s'_i be some R^ψ -deviation of agent i from s , and suppose towards a contradiction that

$$(g(z'), \beta_h(z')) P_i^\psi (g(z), \beta_h(z)), \quad (2)$$

where $z' = H((s'_i, s_{-i})(R^\psi)|_h)$ and $z = H(s(R^\psi)|_h)$. Since $\beta_h(z') = \text{L}_h(z', s) \subseteq \text{LD}_h(z', s)$ by the second restriction, and since s is maximally-dispersed or outcome-condensed it follows that $\beta_h(z') = \{R\}$.

Now, since s° is a subgame perfect equilibrium (SPE), it cannot be the case that $z' P_i z$. If it were, then s'_i would be a profitable local deviation from s° at R , which contradicts the assumption that s° is a SPE.

Thus, it must be the case that zR_iz' . If s is maximally-dispersed, and so $\beta_h(z) = \{R\}$, then zR_iz' implies that

$$(g(z), \{R\}) R_i^\psi (g(z'), \{R\}).$$

Thus,

$$(g(z), \beta_h(z, s)) R_i^\psi (g(z'), \beta_h(z', s)),$$

contradicting (2) above.

Alternatively, if s is maximally-condensed, and so $\beta_h(z) = \mathcal{R}|_a$ and $z = z' = a$, then since ψ is privacy favoring it must be the case that

$$(a, \mathcal{R}|_a) R_i^\psi (a, \{R\}).$$

Thus, once again

$$(g(z), \beta_h(z)) R_i^\psi (g(z'), \beta_h(z')),$$

contradicting (2) above.

Thus, there can be no beneficial R^ψ -deviation of agent i at h , and so s is an information-sensitive subgame perfect equilibrium of (H, A, g) at ψ . \blacksquare

Lemma C.5 *In a mechanism (H, A, g) , if (s, β) is an information-sensitive subgame perfect equilibrium at some lexicographic Ψ for any β , then the profile s° satisfying $s_i^\circ(R, h) = s_i(R^\psi, h)$ for all i, R , and h is a subgame perfect equilibrium of (H, A, g) .*

Proof: Suppose towards a contradiction s° is not a SPE. This implies that there exists an agent $i \in N$, a history $h \in H \setminus Z$, a profile $R \in \mathcal{R}$, and an R -deviation s'_i of agent i such that $g(z')P_i g(z)$, where $z' = H(s'_i(R), s_{-i}^\circ)|_h$ and $z = H(s^\circ(R))|_h$.

Fix some $\psi \in \Psi$. Let s''_i be the R^ψ -deviation of agent i from s that satisfies $s''_i(R^\psi, h) = s'_i(R, h)$ for all R and h , and observe that

$$z'' = H((s'_i, s_{-i})'(R^\psi)|_h)$$

and

$$z = H(s(R^\psi)|_h).$$

Thus, $g(z'')P_i^\psi g(z)$. Setting $S = \beta_h(z'')$ and $T = \beta_h(z)$, we get that

$$(g(z''), \beta_h(z'')) P_i^\psi (g(z), \beta_h(z)),$$

contradicting the assumption that s is an information-sensitive subgame perfect equilibrium at ψ . Hence, s^o is a SPE of (H, A, g) . \blacksquare

We also have a similar lemma for the case of MWR preferences, but first need a definition.

Definition C.6 (maximal deviability) *A mechanism (H, A, g) satisfies maximal deviability with respect to an SCC F if the following holds for any $R \in \mathcal{R}$, profile s satisfying $g(H(s(R))) \notin F(R)$, and history h : Under R , either s is a Nash equilibrium of the subgame rooted at h , or some agent i has a deviation s'_i that will yield an outcome in $\text{TR}_i(R)$.*

Lemma C.7 *Fix a mechanism (H, A, g) that satisfies maximal deviability with respect to an SCC F , and suppose the pair (s, β) is an information-sensitive subgame perfect equilibrium at some MWR Ψ for any β . Then for every $R \in \mathcal{R}$ and $\psi \in \Psi$, either $g(H(s(R^\psi))) \in F(R)$ or the profile $s^o(R)$ satisfying $s^o_i(R, h) = s_i(R^\psi, h)$ for all i and h is a subgame perfect equilibrium of (H, A, g) at R .*

Proof: Suppose towards a contradiction that for some $R \in \mathcal{R}$ it holds that both $g(H(s(R^\psi))) \notin F(R)$ and that $s^o(R)$ is not a SPE at R . Observe that $g(H(s(R^\psi))) = g(H(s^o(R)))$, and so $g(H(s^o(R))) \notin F(R)$. This implies that there exists an agent $i \in N$, a history $h \in H \setminus Z$, and an R -deviation s'_i of agent i such that $z' P_i z$, where $z' = g(H((s'_i, s^o_{-i})(R))|_h)$ and $z = g(H(s^o(R))|_h)$. Since (H, A, g) satisfies maximal deviability with respect to F and $g(H(s^o(R))) \notin F(R)$, we can choose i , h , and s'_i in such a way that $z' \in \text{TR}_i(R)$.

Let s''_i be the R^ψ -deviation of agent i from s that satisfies $s''_i(R^\psi, h) = s'_i(R, h)$ for all R and h , and observe that

$$z'' = H((s'_i, s^o_{-i})'(R^\psi)|_h)$$

and

$$z = H(s(R^\psi)|_h).$$

Thus, $g(z'') P_i^\psi g(z)$ and $g(z'') \in \text{TR}_i(R)$. Furthermore, since ψ satisfies MWR it also holds that

$$(g(z''), S) P_i^\psi (g(z), T)$$

for any sets S and T with $R \in T$. Setting $S = \beta_h(z'')$ and $T = \beta_h(z)$, we get that

$$(g(z''), \beta_h(z'')) P_i^\psi (g(z), \beta_h(z)),$$

contradicting the assumption that s is an information-sensitive subgame perfect equilibrium at ψ . Hence, s^o is a SPE of (H, A, g) . \blacksquare

Lemma C.8 *Fix some Ψ and suppose the number of agents $N \geq 3$. Then the strategy profile s for which $s(R^\psi, h) = s^{\text{MR}'}(R, h)$ for all R, ψ , and h is maximally-dispersed at Ψ in the MR' mechanism.*

Proof: Fix some $i \in N$, $R \in \mathcal{R}$, $\psi \in \Psi$, and $h \in H \setminus Z$, and let $z = H(s(R^\psi) | h)$. Observe that, since the number of agents is at least 3, and since all agents submit the same actions at every history, it is the case that $L_h(z, s) = \text{LD}_h(z, s)$. Furthermore, since there is exactly one $\bar{R} \in \mathcal{R}$ such that $H(s(\bar{R}^\psi) | h) = z$ (namely, $\bar{R} = R$), it holds that $\text{LD}_h(z, s) = L_h(z, s) = \{R\}$.

Next, observe that when a single agent deviates from s at (R, h) to yield a terminal history z' , it is always possible to uniquely determine R from h, s , and z' . This follows from the facts that $L_h(z, s) = \{R\}$, that the number of agents is at least three, and that agents always submit the same actions. Thus, for any i and R^ψ -deviation $s'_i \neq s_i$ of agent i it holds that

$$\text{LD}_h(H((s'_i, s_{-i})(R^\psi) | h), s) = \{R\},$$

and so s is maximally-dispersed. \blacksquare

C.2 Proofs of Propositions 5.1 and 5.4

Proof of Proposition 5.1: By Theorem 1 of Abreu and Sen (1990), if an SCF f is subgame perfect implementable then f satisfies Condition α . Furthermore, by Theorem 2 of Abreu and Sen (1990), since f also satisfies NVP, it is implementable via the Moore-Repullo mechanism MR. It is then also implementable by mechanism MR' , since the enlarged set of actions there has no effect on the outcomes of the mechanism.

Now, consider the strategy profile s^* for which $s^*(R^\psi, h) = s^{\text{MR}'}(R, h)$ for all R, ψ , and h . By Lemma C.8, this strategy profile is maximally-dispersed at Ψ . By

Lemma C.4, s^* is an information-sensitive subgame perfect equilibrium of MR' . In addition, by the properties of this strategy profile, $s^*(R^\psi) = f(R)$ for all $R \in \mathcal{R}$ and $\psi \in \Psi$. Thus, bullets 1(a) and 1(b) of Definition 2.6 are satisfied.

Furthermore, for any information-sensitive subgame perfect equilibrium s in MR' at $\psi \in \Psi$, Lemma C.5 implies that there is a corresponding SPE with respect to o (since ψ is lexicographic by assumption). However, since MR' is a subgame perfect implementation of f , it must then be the case that $g(H(s(R))) = f(R)$ for all $R \in \mathcal{R}$. Since $H(s^*(R^\psi)) = H(s(R))$ it holds that $g(H(s(R^\psi))) = f(R)$ for all $R \in \mathcal{R}$, satisfying bullet 2 of Definition 2.6.

Thus, MR' is an information-sensitive implementation of the SCF f with respect to any lexicographic ψ . ■

Proof of Proposition 5.4: The mechanism we will use is MR' . Consider the strategy profile s^* for which $s^*(R^\psi, h) = s^{\text{MR}'}(R, h)$ for all R and h . By Lemma C.8, this strategy profile is maximally-dispersed at Ψ . By Lemma C.4, s is an information-sensitive subgame perfect equilibrium of MR' . In addition, by the properties of this strategy profile, $s^*(R^\psi) = f(R)$ for all $R \in \mathcal{R}$ and $\psi \in \Psi$. Thus, bullets 1(a) and 1(b) of Definition 2.6 are satisfied.

Furthermore, Lemma C.9 below states that for any SCC F satisfying Condition α^{\max} and NVP, the Moore-Repullo mechanism, and hence also the mechanism MR' , satisfy maximal deviability with respect to F . So for any information-sensitive subgame perfect equilibrium s in MR' with respect to $\psi \in \Psi$ and any $R \in \mathcal{R}$, Lemma C.7 implies that either $g(H(s(R^\psi))) = f(R)$, or there is a corresponding SPE s^o at R with respect to o . However, since MR' is a subgame perfect implementation of f , it must be the case that $g(H(s^o(R))) = f(R)$, and so also $g(H(s^o(R))) = f(R)$. Since $H(s^*(R^\psi)) = H(s(R))$ it holds that $g(H(s(R^\psi))) = f(R)$ for all $R \in \mathcal{R}$ and $\psi \in \Psi$, satisfying bullet 2 of Definition 2.6.

Thus, MR' is an information-sensitive implementation of the SCF f with respect to any MWR ψ . ■

Lemma C.9 *Under NVP, the implementation of an SCC F satisfying Condition α^{\max} using the Moore-Repullo mechanism satisfies maximal deviability with respect to F .*

Proof: The proof follows Theorem 2 of Abreu and Sen (1990), with the observation that whenever an agent has a deviation from some strategy profile s that does not lead to an outcome in F , he actually has a deviation that will yield an outcome that is top-ranked by him. ■

C.3 Proofs of Theorems 5.2 and 5.5

The proofs of Theorems 5.2 and 5.5 both use the following mechanism MRP:

Mechanism MRP:

- **Stage 0(a):** Each agent i simultaneously submits a pair $(a_i, n_i) \in \mathcal{O} \times \mathbb{Z}$. If $a_1 = \dots = a_N$ then the outcome is $g(a_1, \dots, a_N) = a_1$. If there exists $j_d \in N$ and $a \in \mathcal{O}$ such that $a_i = a$ for all agents $i \neq j_d$ but $a_{j_d} \neq a$, then go to Stage 0(b). Otherwise, the agent who submitted the highest integer n_i chooses the outcome of the mechanism.
- **Stage 0(b):** Suppose the history is $h = ((a_1, n_1), \dots, (a_N, n_N))$. Each agent i simultaneously submits a pair $(R^i, n^i) \in \{\mathcal{R} \cup \{\perp\}\} \times \mathbb{Z}$.
 - (i) If there does *not* exist an agent j and a profile $R \in \mathcal{R}$ such that $R^i = R$ for all $i \in N \setminus \{j\}$, then the agent who submitted the highest integer n^i chooses the outcome of the mechanism.
 - (ii) If there *does* exist an agent j and a profile $R \in \mathcal{R}$ such that $R^i = R$ for all $i \in N \setminus \{j\}$ and if $a_i = f(R)$ for all $i \in N \setminus \{j, j_d\}$ then the outcome is $f(R)$, unless agent j announces R^j with $f(R^j) \neq f(R)$ and $j = j(0)$ in the sequence $j(R, R^j, f(R))$. In this latter case:
 1. Fix $a^i \stackrel{\text{def}}{=} a_i$ for all $i \in N \setminus \{j_d\}$;
 2. If $j \neq j_d$ then fix $a^{j_d} \stackrel{\text{def}}{=} a$; If $j = j_d$ then fix $a^{j_d} \stackrel{\text{def}}{=} a_{j_d}$.
 3. Go to Stage 1.
 - (iii) In all other cases, agent j_d chooses the outcome of the mechanism.
- **Stage k , $k = 1, \dots, \ell$:** Continue with Stage k of the MR' mechanism, where the Stage 0 history is $((R^1, a^1, n^1), \dots, (R^N, a^N, n^N))$.

Consider the following strategy s^{MRP} :

- In Stage 0(a), $s_i^{\text{MRP}}(R, \epsilon) = (f(R), 0)$ for every $i \in N$.
- In Stage 0(b), profile R , and history h , every agent i submits $s_i^{\text{MRP}}(R, h) = (R, 0)$.
- For all subsequent rounds (of the MR' part of the mechanism), agents always announce $(0, R)$, as in $s^{\text{MR}'}$.

We now proceed with various lemmas that will be used in the proofs of Theorems 5.2 and 5.5.

Lemma C.10 *If f satisfies α , then for every R the profile s^{MRP} is a subgame perfect equilibrium of MRP at R .*

Proof: First observe that s^{MRP} is an SPE in Stages 1 and onwards: At history $((a_1, n_1), \dots, (a_N, n_N), (R^1, n^1), \dots, (R^N, n^N))$ and onwards, the strategy and mechanism are identical to the SPE strategy $s^{\text{MR}'}$ in MR' at history $((R^1, a^1, n^1), \dots, (R^N, a^N, n^N))$ and onwards.

Now consider Stage 0(b) and a history $h = ((a_1, n_1), \dots, (a_N, n_N))$. If there exists $j_d \in N$ such that $a_i = f(R)$ for all agents $i \neq j_d$, then the agents are essentially playing Stage 0 of MR' (this is case (i)). Thus, since s^o is identical to $s^{\text{MR}'}$ here, there is no profitable deviation. On the other hand, if there does not exist $j_d \in N$ such that $a_i = f(R)$ for all agents $i \neq j_d$, then, regardless of any unilateral deviation, s^{MRP} will lead to case (ii), and so the outcome will be chosen by agent j_d . To see this, observe that the history must be such that there exists some agent j_d and an outcome $b \neq f(R)$ such that $a_i = b$ for all $i \neq j_d$. If no agent deviates, then the outcome will be chosen by agent j_d , since it is not the case that $a_i = f(R)$ for any i , and so this leads to case (ii). If some agent does deviate, then the mechanism can still determine the true R from agents' messages (since $N - 1$ agents agree on the same R). However, even then it will not be the case that $a_i = f(R)$ for any agent i except possibly j_d . This, a unilateral deviation also leads to case (ii).

Finally, consider Stage 0(a) of the mechanism. Observe that the strategy profile s^{MRP} here is identical to $s^{\text{MR}'}$. Thus, any profitable unilateral deviation here would

imply a profitable deviation from $s^{\text{MR}'}$ in MR' . Since $s^{\text{MR}'}$ is an SPE profile, there is no profitable deviation from s^{MRP} in Stage 1 either. \blacksquare

Lemma C.11 *If f satisfies α and NVP, then for any $R \in \mathcal{R}$ all subgame perfect equilibria of MRP at R lead to the outcome $f(R)$.*

Proof: Suppose towards a contradiction that there is a SPE s of (H, A, g) that, at R , leads to an outcome $b \neq f(R)$. Observe first that in Stage 0(a) of the mechanism it must be the case that $s_i(R, \epsilon) = b$ for all $i \in N$. Otherwise, if not all agents agree on the same outcome b , then at least $N - 1$ agents have a unilateral deviation that would yield them a top-ranked outcome. Thus, b must be top-ranked by all these agents. But if b is a top-ranked outcome for $N - 1$ agents, then by NVP it must be the case that $b = f(R)$, a contradiction.

Now consider Stage 0(b) with a history $h = ((a_1, n_1), \dots, (a_N, n_N))$. Suppose that for every $i \in N$ the strategy is $s(R, h) = (R^i, n^i)$. There are two cases to consider:

1. There exists an $\bar{R} \in \mathcal{R}$ such that $R^i = \bar{R}$ for all i , and $b = f(\bar{R})$. In this case, we claim that agent $j(0)$ in the sequence $j(\bar{R}, R, b)$ has a profitable deviation that yields him a top-ranked outcome. In particular, he can deviate to $s_{j(0)}(h, R) = (R, 0)$. This leads to case (ii) of Stage 0(b), and a continuation to Stage 1. This situation is then the same as the subgame of MR' that follows the history h' in which all agents but $j(0)$ announce (\bar{R}, b) and agent $j(0)$ announces (R, a) . By Lemma C.1, all subgame perfect equilibrium outcomes of this subgame are top-ranked by agent $j(0)$. Thus, since agent $j(0)$ has a profitable deviation leading to a top-ranked outcome, this case is not an equilibrium.
2. There exists an $\bar{R} \in \mathcal{R}$ and $j \in N$ such that $R^i = \bar{R}$ for all $i \in N \setminus \{j\}$, and $b = f(\bar{R})$. There are two sub-cases to consider:
 - (a) $j \neq j_d$. In this case, agent j_d has a deviation to trigger (and win) the integer game of case (i), by submitting $R^{j_d} = \perp$. Thus, either this case is not an equilibrium, or j_d already gets a top-ranked outcome.
 - (b) $j = j_d$. If $j_d \neq j(0)$, then agent $j(0)$ has a profitable deviation yielding him a top-ranked outcome – namely, he triggers the integer game of case (i) by submitting $R^{j(0)} = \perp$. This deviation is profitable by (iii) of Condition

α . Alternatively, if $j_d = j(0)$ then he can deviate to $s_{j(0)}(h, R) = (R, 0)$. This leads to case (ii) of Stage 0(b), and a continuation to Stage 1. As in case 1 above, this situation is the same as the subgame of MR' that follows the history h' in which all agents but $j(0)$ announce (\bar{R}, b) and agent $j(0)$ announce (R, a) . Again, by Lemma C.1, all subgame perfect equilibrium outcomes of this subgame are top-ranked by agent $j(0)$. Thus, since agent $j(0)$ has a profitable deviation leading to a top-ranked outcome, this case is not an equilibrium.

Thus, in all cases above, the equilibrium outcomes are always top-ranked by agent j_d . All other cases not included above are captured by case (iii) in Stage 0(b) of the protocol, in which j_d chooses an outcome. Thus, whenever the equilibrium profile $s(R)$ leads to an outcome $b \neq f(R)$, a deviation by an agent j_d would lead to a subgame in which all SPE yield an outcome that is top-ranked by j_d . In particular, this means that there can be no such equilibrium s : Since $s(R)$ leads to an outcome $b \neq f(R)$, there must exist an agent j_d for whom b is not top-ranked (by NVP). Thus, j_d will deviate, leading to an outcome that is top-ranked by him. ■

Lemma C.12 s^{MRP} is maximally-dispersed.

Proof: Fix some $i \in N$, $R \in \mathcal{R}$, $h \in H \setminus Z$, and R^ψ -deviation $s'_i \neq s_i$ of agent i . Since the prescription of s^{MRP} is for every agent to play the same strategy at h and $N \geq 3$, when an agent deviates this deviation is noticeable. Furthermore, there is no \bar{R}^ψ that leads to the same outcome under this deviation: That is, for every $\bar{R}^\psi \neq R^\psi$ it holds that $H((s'_i, s_{-i})(\bar{R}^\psi) | h) \neq H((s'_i, s_{-i})(R^\psi) | h)$. Thus,

$$\text{LD}_h(H((s'_i, s_{-i})(R^\psi) | h), s) = \{R\}.$$

Now, in Stage 0(a), it holds that

$$L(H(s(R^\psi)), s) = \text{LD}(H(s(R^\psi)), s) = \mathcal{R}|_a.$$

Finally, in Stages 0(b) and onwards agents always submit the action R as part of their strategies. No deviation from any $R' \neq R$ will lead to the same coordination on R . Thus,

$$\text{LD}_h(H(s(R^\psi) | h), s) = \{R\}.$$

■

Lemma C.13 *If f satisfies α^{\max} and NVP, then MRP satisfies maximal deviability with respect to f .*

Proof: Fix some $R \in \mathcal{R}$ and a strategy profile s for which $g^{\text{MRP}}(H(s(R))) = b \neq f(R)$. Suppose first that in Stage 0(a) of the mechanism it is *not* the case that $s_i(R, \epsilon) = b$ for all $i \in N$. Thus, at least $N - 1$ agents have a unilateral deviation that would trigger the integer game and yield them a top-ranked outcome. By NVP and the fact that $b \neq f(R)$, it must be the case that for at least one of these $N - 1$ agents b is not top-ranked. Thus, this agent has a profitable deviation that yields him a top-ranked outcome.

Suppose next that in Stage 0(a) of the mechanism it holds that $s_i(R, \epsilon) = b$ for all $i \in N$, and consider Stage 0(b) with a history $h = ((a_1, n_1), \dots, (a_N, n_N))$. Suppose that for every $i \in N$ the strategy is $s(R, h) = (R^i, n^i)$. There are two cases to consider:

1. There exists an $\bar{R} \in \mathcal{R}$ such that $R^i = \bar{R}$ for all i , and $b = f(\bar{R})$. In this case, consider a deviation by agent $j(0)$ in the sequence $j(\bar{R}, R, b)$ to $s_{j(0)}(h, R) = (R, 0)$. This leads to case (ii) of Stage 0(b), and a continuation to Stage 1. This situation is then the same as the subgame of MR' that follows the history h' in which all agents but $j(0)$ announce (\bar{R}, b) and agent $j(0)$ announces (R, a) . By Lemma C.9, either $s(R)$ is a SPE of the subgame rooted at h' , or some agent has a deviation that yields him a top-ranked outcome. In the latter case, we are done. For the former case, Lemma C.1 implies that all SPE outcomes of this subgame are top-ranked by agent $j(0)$. Thus, agent $j(0)$ has a profitable deviation at Stage 0(b) that leads to a top-ranked outcome.
2. There exists an $\bar{R} \in \mathcal{R}$ and $j \in N$ such that $R^i = \bar{R}$ for all $i \in N \setminus \{j\}$, and $b = f(\bar{R})$. There are two sub-cases to consider:
 - (a) $j \neq j_d$. In this case, agent j_d has a deviation to trigger (and win) the integer game of case (i), by submitting $R^{j_d} = \perp$. Thus, either agent j_d has a profitable deviation yielding him a top-ranked outcome, or j_d already gets a top-ranked outcome.
 - (b) $j = j_d$. If $j_d \neq j(0)$, then agent $j(0)$ has a profitable deviation yielding him a top-ranked outcome – namely, he triggers the integer game of case (i)

by submitting $R^{j(0)} = \perp$. This deviation is profitable by (iii) of Condition α^{\max} . Alternatively, if $j_d = j(0)$ then he can deviate to $s_{j(0)}(h, R) = (R, 0)$. This leads to case (ii) of Stage 0(b), and a continuation to Stage 1. As in case 1 above, this situation is the same as the subgame of MR' that follows the history h' in which all agents but $j(0)$ announce (\bar{R}, b) and agent $j(0)$ announce (R, a) . Again, either $s(R)$ is a SPE of the subgame rooted at h' , in which case agent $j(0)$ has a profitable deviation at Stage 0(b) that leads to a top-ranked outcome, or some agent has a deviation that yields him a top-ranked outcome.

Thus, in all cases above, either some agent has a deviation from $s(R)$ that leads to a top-ranked outcome, or the outcome following a deviation by agent j_d is top-ranked by agent j_d . All other cases not included above are captured by case (iii) in Stage 0(b) of the protocol, in which j_d chooses an outcome. Thus, whenever the profile $s(R)$ leads to an outcome $b \neq f(R)$, either some agent has a deviation yielding him a top-ranked outcome, or a deviation by an agent j_d would lead to a subgame in which all outcomes are top-ranked by j_d . In particular, this means that, since $s(R)$ leads to an outcome $b \neq f(R)$, there must exist an agent j_d for whom b is not top-ranked (by NVP). Thus, j_d has a profitable deviation leading to an outcome that is top-ranked by him. ■

We are now ready to prove Theorems 5.2 and 5.5.

Proof of Theorem 5.2: There are two parts to the proof, corresponding to the two bullets of Definition 2.7. The first part of the proof follows. Since f can be implemented in subgame perfect equilibrium, by Abreu and Sen (1990) f must satisfy Condition α . By Lemma C.10, for every $R \in \mathcal{R}$ the strategy $s^{\text{MRP}}(R)$ is a SPE of MRP at R . Furthermore, by Lemma C.12, s^{MRP} is maximally-dispersed. Thus, by Lemma C.4, the strategy profile s^* satisfying $s_i^*(R^\psi, h) = s_i^{\text{MRP}}(R, h)$ for all i, R , and h is an information-sensitive subgame perfect equilibrium at ψ . Note that for all $R, \bar{R} \in \mathcal{R}$ such that $f(R) = f(\bar{R})$ it holds that $s^*(R^\psi, \epsilon) = s^*(\bar{R}^\psi, \epsilon)$, and that these both lead to terminal histories. In particular, this implies that $H(s^*(\mathcal{R}^\psi)) = H(s^*(\bar{\mathcal{R}}^\psi))$, and so the strategy profile s^* satisfies bullets 1(a) and 1(b) of Definition 2.6 and bullet 1(c) of Definition 2.7.

Next, the second part of the proof is as follows. Let s be an information-sensitive subgame perfect equilibrium of MRP at ψ with respect to some β . Since ψ is lexicographic, Lemma C.5 implies that the profile s^o satisfying $s_i^o(R, h) = s_i(R^\psi, h)$ for all i , R , and h is a SPE of MRP. Now, since f satisfies NVP by assumption, Lemma C.11 implies that for every $R \in \mathcal{R}$, all SPE of MRP at R lead to the outcome $f(R)$. Since s and s^o lead to the same outcome, this implies that for every $R \in \mathcal{R}$, the profile $s(R^\psi)$ leads to the outcome $f(R)$. Thus, bullet 2 of Definition 2.6 is also satisfied.

Hence, the mechanism MRP is a privacy-protecting implementation of f at ψ . ■

Proof of Theorem 5.5: As in the proof of Theorem 5.2 there are two parts to the proof. The first part is identical to the proof of Theorem 5.2, implying the existence of a strategy profile s^* satisfying bullets 1(a) and 1(b) of Definition 2.6 and bullet 1(c) of Definition 2.7.

The second part of the proof is as follows. First note that, by Lemma C.13, the mechanism MRP satisfies maximal deviability with respect to f . Let s be an information-sensitive subgame perfect equilibrium of MRP at ψ with respect to some β . Since ψ satisfies MWR, Lemma C.7 implies that, for every $R \in \mathcal{R}$, the profile s^o satisfying $s_i^o(R, h) = s_i(R^\psi, h)$ for all i and h either leads to the outcome $f(R)$, or it is a SPE of MRP at R . In the former case, since $s(R^\psi)$ and $s^o(R)$ lead to the same outcome, $s(R^\psi)$ leads to the outcome $f(R)$. For the latter case recall that f satisfies NVP by assumption, and so Lemma C.11 implies that for every $R \in \mathcal{R}$, all SPE of MRP at R lead to the outcome $f(R)$. Again, since s and s^o lead to the same outcome, this implies that for every $R \in \mathcal{R}$, the profile $s(R^\psi)$ leads to the outcome $f(R)$. Thus, bullet 2 of Definition 2.6 is also satisfied.

Hence, the mechanism MRP is a privacy-protecting implementation of f at ψ . ■

D Proofs from Section 6

D.1 Proofs of Theorems 6.7 and 6.8

The mechanism used for Theorems 6.7 and 6.8 is the following variant of MRP from Section C.3. The mechanism is parametrized by a partition Π of \mathcal{R} . Label each element of Π by a unique number, and denote by $\pi(R)$ the label given to the element

of Π that contains R .

Mechanism MRP^Π :

- **Stage 0(a):** Each agent i simultaneously submits a pair $(a_i, n_i, \pi_i) \in \mathcal{O} \times \mathbb{Z} \times \mathbb{N}$. If $a_1 = \dots = a_N$, $\pi_1 = \dots = \pi_N$, and $f(R) = a_1$ for all R satisfying $\pi(R) = \pi_1$, then the outcome is a_1 . If there exists $j_d \in N$, $a \in \mathcal{O}$, and $\pi \in \mathbb{N}$ such that $a_i = a$ and $\pi_i = \pi$ for all agents $i \neq j_d$ but $a_{j_d} \neq a$, then go to Stage 0(b). Otherwise, the agent who submitted the highest integer n_i chooses the outcome of the mechanism.
- **Stage 0(b):** Suppose the history is $h = ((a_1, n_1, \pi_1), \dots, (a_N, n_N, \pi_N))$. Continue as in MRP.
- **Stage k , $k = 1, \dots, \ell$:** Continue as in MRP.

Consider the following strategy s^{MRP^Π} :

- In Stage 0(a), $s_i^{\text{MRP}^\Pi}(R, \epsilon) = (f(R), 0, \pi(R))$ for every $i \in N$.
- In Stage 0(b), profile R , and history h , every agent i submits $s_i^{\text{MRP}^\Pi}(R, h) = (R, 0)$.
- For all subsequent rounds (of the MR' part of the mechanism), agents always announce $(0, R)$, as in $s^{\text{MR}'}$.

The proofs of Theorems 6.7 and 6.8 are nearly the same as those of Theorems 5.2 and 5.5. The only difference is in condition 1(c) of Definition 6.4. We need to show that if R and \bar{R} are such that $\bar{R} \in \Pi(R)$, then $H(s^{\text{MRP}^\Pi}(\mathcal{R}^\psi)) = H(s^{\text{MRP}^\Pi}(\bar{\mathcal{R}}^\psi))$. We also need to show that if R and \bar{R} are such that $\bar{R} \notin \Pi(R)$, then $H(s^{\text{MRP}^\Pi}(\mathcal{R}^\psi)) \neq H(s^{\text{MRP}^\Pi}(\bar{\mathcal{R}}^\psi))$. These are immediate from the specification of the strategy profile s^{MRP^Π} : In the former case, the profile prescribes the same strategy, since in that case $\pi(R) = \pi(\bar{R})$. In the latter case the prescription is different, since in that case $\pi(R) \neq \pi(\bar{R})$ (and so the history differs in Stage 0(a)).

References

- ABREU, D. and SEN, A. (1990). Subgame perfect implementation: A necessary and almost sufficient condition. *Journal of Economic theory*, **50** 285–299.
- ABREU, D. and SEN, A. (1991). Virtual implementation in nash equilibrium. *Econometrica: Journal of the Econometric Society* 997–1021.
- ACQUISTI, A., TAYLOR, C. and WAGMAN, L. (2016). The economics of privacy. *Journal of Economic Literature*, **54** 442–492.
- AUMANN, R. J. (1976). Agreeing to disagree. *The annals of statistics* 1236–1239.
- BERNHEIM, B. (1994). A theory of conformity. *Journal of political Economy* 841–877.
- CHEN, J. and MICALI, S. (2015). Mechanism design with possibilistic beliefs. *Journal of Economic Theory*, **156** 77–102.
- CHEN, Y., CHONG, S., KASH, I., MORAN, T. and VADHAN, S. (2011). Truthful mechanisms for agents that value privacy. *arXiv preprint arXiv:1111.5472*.
- CREMER, J. and MCLEAN, R. P. (1988). Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society* 1247–1257.
- DINUR, I. and NISSIM, K. (2003). Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 202–210.
- DUTTA, B. and SEN, A. (2012). Nash implementation with partially honest individuals. *Games and Economic Behavior*, **74** 154–169.
- GEANAKOPOLOS, J., PEARCE, D. and STACCHETTI, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, **1** 60–79.
- GHOSH, A. and ROTH, A. (2011). Selling privacy at auction. In *Proceedings of the 12th ACM conference on Electronic commerce*. ACM, 199–208.

- GLAZER, A. and KONRAD, K. (1996). A signaling explanation for charity. *The American Economic Review*, **86** 1019–1028.
- GLAZER, J. and RUBINSTEIN, A. (1998). Motives and implementation: On the design of mechanisms to elicit opinions. *Journal of Economic Theory*, **79** 157–173.
- GRADWOHL, R. (2017). Voting in the limelight. *Economic Theory*, *forthcoming*.
- GRADWOHL, R. and SMORODINSKY, R. (2017). Perception games and privacy. *Games and Economic Behavior*, **104** 293–308.
- IRELAND, N. (1994). On limiting the market for status signals. *Journal of public Economics*, **53** 91–110.
- JACKSON, M. (2001). A crash course in implementation theory. *Social choice and welfare*, **18** 655–708.
- MASKIN, E. (1999). Nash equilibrium and welfare optimality. *The Review of Economic Studies*, **66** 23–38.
- MASKIN, E. and SJÖSTRÖM, T. (2002). Implementation theory. *Handbook of social Choice and Welfare*, **1** 237–288.
- MATSUSHIMA, H. (2008a). Behavioral aspects of implementation theory. *Economics Letters*, **100** 161–164.
- MATSUSHIMA, H. (2008b). Role of honesty in full implementation. *Journal of Economic Theory*, **139** 353–359.
- MCSHERRY, F. and TALWAR, K. (2007). Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*. IEEE Computer Society, 94–103.
- MILTERSEN, P., NIELSEN, J. and TRIANOPOULOS, N. (2009). Privacy-enhancing auctions using rational cryptography. *Advances in Cryptology-CRYPTO 2009* 541–558.
- MOORE, J. and REPULLO, R. (1988). Subgame perfect implementation. *Econometrica: Journal of the Econometric Society* 1191–1220.

- MULLER, E. and SATTERTHWAITE, M. (1977). The equivalence of strong positive association and strategy-proofness. *Journal of Economic Theory*, **14** 412–418.
- NAOR, M., PINKAS, B. and SUMNER, R. (1999). Privacy preserving auctions and mechanism design. In *Proceedings of the 1st ACM conference on Electronic commerce*. ACM, 129–139.
- NISSIM, K., ORLANDI, C. and SMORODINSKY, R. (2012). Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 774–789.
- PALFREY, T. (2002). Implementation theory. *Handbook of Game Theory with Economic Applications*, **3** 2271–2326.
- POSTLEWAITE, A. and WETTSTEIN, D. (1989). Feasible and continuous implementation. *The Review of Economic Studies*, **56** 603–611.
- SAIJO, T. (1987). On constant maskin monotonic social choice functions. *Journal of Economic Theory*, **42** 382–386.
- SERRANO, R. (2004). The theory of implementation of social choice rules. *SIAM Review*, **46** 377–414.
- VARTIAINEN, H. (2007). Subgame perfect implementation: A full characterization. *Journal of Economic Theory*, **133** 111–126.
- VOHRA, R. (2012). Dynamic mechanism design. *Surveys in Operations Research and Management Science*, **17** 60–68.
- XIAO, D. (2011). Is privacy compatible with truthfulness. Tech. rep., Cryptology ePrint Archive, Report 2011/005.