SAFETY IN MECHANISM DESIGN AND IMPLEMENTATION THEORY[†]

GEORGE SHOUKRY[*]

ABSTRACT. We introduce safety in implementation theory as robustness of players' welfare to unexpected deviations by others. Specifically, we define a (α,ω)-safe strategy profile to be one at which a deviation by any α players can result in at most ω victims, and define a victim to be a player who attains an outcome within his least preferred K outcomes, where K is set by the designer. We discuss three notions of implementation: Nash implementation with all Nash equilibria being safe, implementation in safe equilibria, and double implementation in safe and Nash equilibria. Strong safe implementation is the most desirable notion; we show that it is restrictive when no other assumptions are made, but that mild behavioral assumptions unrelated to safety can achieve strong safe implementation from double implementation. We give conditions that are necessary for implementation in safe equilibria and sufficient under mild preference restrictions. We then address double implementation and, allowing for simple transfers, we give conditions that are both necessary and sufficient for double implementation. We prove a "no guarantees" result showing that absolute safety ($\alpha = n-1$ and $\omega = 0$) can lead to non-existence of Nash equilibria in a wide range of environments making safe implementation with guarantees impossible. We show that, in general, safety leads to restrictions in every player's freedom and to a loss in efficiency. We illustrate the design of a safe mechanism in an example of hiring candidates with multidimensional quality.

1 INTRODUCTION

We introduce a notion of safety as robustness of agents' welfare to unexpected deviations by others. Such robustness is crucial in the current global environment, where malicious intentions, seemingly irrational behavior, mistakes, accidents, or miscommunication can all lead to disasters. Without safety constraints, a mechanism leaves participants exposed to arbitrarily harmful outcomes resulting from unexpected deviations by others – arguably unacceptable, especially as mechanism design extends its reach to more applied settings. Despite the considerable importance of safety, it has received very little attention in the mechanism design and implementation theory literature.

The canonical mechanism for Nash implementation (Maskin (1999)) lacks any safety properties. In an example in section 2.1, we will see that, in general, at *any* Nash equilibrium (NE) of Maskin's mechanism, a single player can impose the most undesirable outcome on all

Date: October 2, 2013.

[†]I would like to thank Jason Abrevaya, Justin Leroux, Laurent Mathevet, Maxwell Stinchcombe, and Thomas Wiseman for their valuable comments.

[*]Dept. of Economics, University of Texas, 1 University Station, C3100, Austin, TX 78712, U.S.A.; gshoukry@utexas.edu.

others by deviating. The lack of safety in Maskin's mechanism applies to many mechanisms in the literature.

Our precise understanding of safety, and exactly what aspects of robustness make a mechanism "safe", have important implications for the analysis and interpretation of this paper's results. We highlight three important aspects.

First, we aim to protect players from deviators, *regardless* of the incentives to deviate. A mechanism can sometimes count on players being self interested for safety of others. For example, many roads do not have safeguards against drivers driving on the wrong side of the road because doing so would harm those drivers. In such settings our definition of safety may be too strong. However, there are many situations where a designer may wish to assure safety regardless of the incentives associated with deviating: to protect against actions of players intending to harm others without regard to their own utility, irrational behavior, mistakes, accidents, miscommunication; or in situations where deviations can be very costly as in flying a passenger jet or controlling a nuclear arsenal where one person can hurt many more. In any of those cases we insist on more regulations because incentives against deviating cannot be counted on to provide safety, making it necessary to adopt an idea of safety that focuses instead on the potential impact of deviations. This is precisely what our notion of safety allows us to do.

Second, we consider safety of players from *unexpected* deviations. We let players be entirely unprepared for deviations from equilibrium by others; players may wish to behave differently if they suspected the existence of deviators, as is plausible in many realistic environments. This is in contrast to Eliaz (2002), where all players know that deviations from equilibrium may occur, and all have accurate and identical beliefs about the maximum number of players who may deviate. Players may be prepared for deviations in some problems, and it is interesting to consider robust mechanisms that work well in such settings, as in Eliaz (2002); however, our approach applies to situations where players are unprepared for deviations. For example, a kidney exchange designer may wish to make agents safe from unexpected behavior by others (e.g. intentionally transmitting deadly viruses) without relying on the assumption that all agents would be best-responding to accurate beliefs about deviators if deviations from equilibrium were possible. We also note that, in reality, mechanisms are often stopped and operations are halted if it is suspected that malicious agents who may harm others are present. Additionally, assuming best-responding players is cognitively demanding in the presence of deviators who may deviate in arbitrary ways. Our approach places the bulk of the safety burden on the designer, who aims to protect unprepared players from unexpected, harmful actions by others.

Third, we focus on robustness of players' *welfare* as opposed to the mechanism's final outcome. The distinction is important, because different robustness requirements lead to

different sets of implementable social choice rules. In many applications (e.g. public good problems), the actual outcome of the mechanism is only of secondary importance to the designer and may only matter due to its indirect effect on the welfare of the players. It is then desirable to design mechanisms that have some robustness properties with respect to the welfare of the players directly.

Combining these aspects results in a robust, intuitive notion of safety that can be applied to examine the safety properties in virtually any setting. We define a (α,ω)-safe strategy profile to be one at which a deviation by any α players (the *deviators*) can result in at most ω victims among the non-deviators. We apply our safety concept to the standard implementation problem by defining a *victim* to be a player who attains an outcome within that player's least favorite K outcomes, where K is set by the designer.

One implication of the deviations being unexpected is that we do not count on the players to protect themselves in anyway–they do not expect deviations from equilibrium. So, we must assume that they play a Nash equilibrium (or any other equilibrium notion of choice) without any regard for safety. In this paper we focus on safety in conjunction with the Nash equilibrium as a solution concept. We introduce safety to mechanism design and implementation theory by considering three notions of implementation.

First, we consider a form of implementation requiring all Nash equilibria to be safe. We define strong (α,ω)-safe implementation to be Nash implementation with the added restriction that all Nash equilibria at any preference profile be (α,ω)-safe. Strong (α,ω)-safe implementation is the most attractive form of safe implementation: every desirable social outcome is obtainable via a Nash equilibrium of the mechanism, and every Nash equilibrium is (α,ω)-safe and leads to a desirable social outcome. Behavioral assumptions on the players are weak because simply playing a Nash equilibrium assures safety. However, we show that strong (α,ω)-safe implementation imposes extreme restrictions on social choice rules and preferences that are unlikely to hold in many natural environments. This gives rise to questions about weaker forms of safe implementation that can be achieved, and what conditions or assumptions are needed for strong safe implementation to be possible; we address these questions by considering two other notions of implementation.

Next, we consider a much weaker form of implementation, called (α,ω)-safe implementation, requiring the following two properties at any preference profile: all desirable outcomes can be obtained via Nash equilibria that are safe, and all Nash equilibria that are safe lead to desirable outcomes. We call a Nash equilibrium that is also safe a (α,ω)-safe equilibrium. (α,ω)-Safe implementation is equivalent to (full) implementation in (α,ω)-safe equilibria; it has the drawback that unsafe Nash equilibria may exist and may lead to undesirable outcome. However, safe implementation forms the basis for implementing social choice rules in a safe way, and it is of theoretical importance for several reasons. First, it can inform us

about the existence of social choice rules that are Nash, but not safe, implementable, and safe, but not Nash, implementable. This is important for understanding the trade-offs of safety. Second, by satisfying the minimum requirements that are needed for implementing social choice rules in a safe way, it can be used as a step towards more desirable forms of implementation. We give conditions that are necessary for safe implementation, and sufficient with mild preference restrictions.

Finally, we consider double implementation in both (α,ω)-safe and Nash equilibria. Double implementation in this case is equivalent to safe implementation with the added requirement that all Nash equilibria (including unsafe ones) lead to desirable outcomes. Thus, it lies between strong safe implementation and safe implementation in restrictiveness: it allows for some unsafe Nash equilibria, but it constrains all of them to lead to desirable outcomes. We allow for simple transfers when considering double implementation. We give conditions that are both necessary and sufficient for double implementation when $n > 2$, $\alpha < n/2$, and the designer does not have access to precise cardinal preference information; we show that double implementation brings us very close to strong safe implementation, and achieves strong safe implementation with mild behavioral assumptions that are unrelated to safety.

Besides our main results on implementation, we discuss other interesting aspects of safety. We prove a "no guarantees" result showing that, in a wide range of environments, absolute safety ($\alpha = n-1$ and $\omega = 0$) leads to non-existence of Nash equilibria, making safe implementation with guarantees impossible.

We discuss the interplay between safety and freedom. We distinguish between negative freedom: freedom from harm, and positive freedom: freedom to obtain a different outcome. We measure positive freedom by how many outcomes a player can obtain via a potential deviation at equilibrium. We show that, in general, *every* player will have no more positive freedom in a safe mechanism than in one that disregards safety. We also show that positive freedom becomes more restricted as mechanisms become safer. Although immediate in our framework, this is a topic of deep and fundamental theoretical importance.

We show that, in general, there is a trade-off between safety and efficiency. There are situations where the designer must choose between Nash implementation and safe implementation, and where both are not possible together (double implementation fails). In those problems, if the designer chooses safe implementation, then it must come at the cost of Nash equilibria that lead to outcomes outside the social choice rule. To the extent that social choice rules attempt to select efficient outcomes, undesirable outcomes that are not in the social choice rule will be inefficient.

Throughout the paper, we use as an example the design of a hiring mechanism in which candidates have multidimensional quality and we contrast a safe mechanism with one that ignores safety.

The following subsection reviews the related literature, and the rest of the paper is organized as follows. Section 2 motivates the need for safety using an example and introduces our notion of safety. Sections 3, 4, and 5 discuss strong safe implementation, safe implementation, and double implementation. Section 6 discusses the "no-guarantees" result and the interplay between safety, freedom, and efficiency. Section 7 illustrates the benefits of using safe mechanisms in a setting of hiring with multidimensional quality. Section 8 discusses our safety concept and notion of implementation in more detail. Section 9 discusses further work and concludes.

1.1 Related Literature

The ideas in this paper relate to various ideas in the robust implementation literature and to some literature in game theory that considers robustness of equilibrium concepts.

Conceptually, there are two strands of robust implementation theory closely related to this paper. However, the robust implementation literature almost exclusively focuses on robustness of the outcome of the mechanisms, where, as mentioned previously, we focus on robustness of players' welfare: (1) Eliaz (2002) defines robustness as the inability of a deviating minority to affect the outcome of the game used in implementation; in our approach, mechanisms do not have to be robust in this sense; rather, deviators can change the outcome of the mechanism, but we want to avoid them causing "disasters" for others. As previously mentioned, another essential feature distinguishing this paper from Eliaz (2002) is our focus on unexpected deviations. (2) Bergemann and Morris (2005) define robustness as insensitivity of the outcome of the mechanism to informational assumptions such as common knowledge, thus, they focus on a different concept of robustness and on the outcome of the mechanism. This paper is the first in the implementation literature to consider robustness of players' welfare as opposed to robustness of outcomes.

In the game theory and computer science literature, some equilibrium concepts consider various forms of robustness to deviations as part of the solution concept. Halpren (2008) defines a k-resilient Nash equilibrium to be a Nash equilibrium where deviations by coalitions of up to k players do not gain by deviating. This is related to the literature on strong Nash equilibria and coalition proof Nash equilibria. Aumann (1960) defines a strong Nash equilibrium (NE) to be a strategy profile where there are no multilateral profitable deviations. In Aumann's definition, the deviations need not be resistant to further multilateral deviations. Bernheim, Peleg, and Whinston (1987) argue that Aumann's definition is too strong. They define a coalition-proof NE to be a strategy profile where there are no multilateral profitable deviations that are also resistant to further deviations by "subcoalitions". This is weaker than Aumann's definition because the set of deviations to be considered is smaller. These

solution concepts focus only on incentives to deviate, whereas we focus on consequences of deviations from equilibrium.

Some equilibrium concepts that do discuss consequences of deviations from equilibrium do so from a *similarity* motivation: the idea that players should behave similarly in similar situations. Such refinements include Selten's (1975) trembling hand perfect equilibrium and Wen-tsun and Jia-he's (1962) essential equilibrium. The trembling hand perfect and essential equilibria require the existence of a NE close to the original one if the strategies of the players and the payoffs, respectively, are perturbed slightly. The similarity motivation for refinements focuses on the effect of deviations from a NE on the behavior of the players, but does not explicitly address the impact of deviations by some on utilities of others. Another notion of equilibrium in the literature that attempts to capture some of the effect of players on each other is risk dominance (Harsanyi and Selten (1988)). Risk dominance selects equilibria based on the willingness of rational players to deviate if they take the riskiness of different equilibria into account. It is based on the idea that players best-respond to beliefs about deviations by others, whereas we consider situations in which players are unprepared for deviations from equilibrium.

Halpren (2008) discusses a robustness property called t-immunity, where a profile is t-immune if no player who does not deviate is worse off if up to t players deviate. This robustness property is very similar to the more general robustness concept we consider in this paper[1], and its use in computer science supports our view that such robustness is important to consider in applications.

Game theoretic equilibrium concepts are ultimately concerned with prediction, but our objective is different: it is to consider when and how it is possible to design mechanisms with desirable robustness properties. The differing objectives imply that desirable aspects of robustness, and the motivations for studying them, may also differ, but this paper's focus on robustness to deviations from equilibrium is generally supported by similar concerns in game theory.

[1]Note that t-immunity sets $\omega = 0$ and uses a "relative" notion of victimization; see our discussion about relative notions of victimization in section 8.

2 Motivating Example and a Notion of Safety

This section motivates the need for safety in mechanisms and presents our safety concept.

2.1 Motivating Example: Hiring with Multidimensional Quality

Consider a board of directors with $n \geq 4$ members, $I = \{1,\ldots,n\}$, where each member $j \in I$ has a different background, $b_j \in \{b^1,\ldots,b^n\}$. The board is deciding on hiring a CEO from among x candidates, $A = \{c_1,\ldots,c_x\}$. Candidates have skill levels s_{c_i} drawn from a continuous, atomless, distribution with support on $[0,m]$, and potentially different backgrounds, $b_{c_i} \in \{b^1,\ldots,b^n\}$, $i \in I$. We assume that there is at least one candidate in each with each of the backgrounds.

The members' preferences are as follows: each member $j \in \{1,\ldots,n\}$ prefers more skilled candidates, but gives a "bonus" premium p_j to candidates with a matching background. Hence, a member j sorts the candidates according to a score for each candidate c_i given by $s_{c_i} + p_j \mathbf{1}\{b_{c_i} = b_j\}$.

We assume that there are no ties in the skill levels of the candidates or the scores given to the candidates by any member so that all preferences of the members are strict.

The desirable social choice function (SCF), F, selects the candidate with the highest skill level. Hence, any mechanism that implements F must always select the highest skilled candidate.[2] Consider Maskin's mechanism (Maskin (1999)). The SCF satisfies no veto power and Maskin monotonicity. In this mechanism each member's strategy space is $A \times \mathscr{R} \times \mathbb{N}$, where $\mathscr{R}$ is the set of possible preference profiles. Let (a^i, R^i, z^i) denote a message reported by member $i \in I$. Given a message profile, $m = (m_1,\ldots,m_n)$, the mechanism selects the outcome $g(m)$ as follows:

Rule 1: If m is such that all members agree on a report, say (a,R,z), and if $a \in F(R)$ then $g(m) = a$.

Rule 2: Suppose all members agree on (a,R,z) with $a \in F(R)$, except for one member i who reports $(a^i,R^i,z^i) \neq (a,R,z)$. If a^i is weakly less preferred to a for member i under R_i then $g(m) = a^i$; otherwise, $g(m) = a$.

Rule 3: In all other cases, let $g(m) = a^{q^*}$ for $q^* = \max\{q \in I : z^q = \max_j z^j\}$.

This mechanism has virtually no safety properties: there are always $n-1$ members who can *each* get the mechanism to select the least skilled candidate by deviating alone from *any*

[2] It must then be possible to deduce which candidate is the highest skilled from a given preference profile of the members. In this environment, a candidate c_i is more skilled than candidate c_j if and only if at least $n-1$ members rank candidate c_i higher than candidate c_j: if c_i is more skilled than candidate c_j then all members, except possibly for the member with the same background as c_j, will rank c_i higher than c_j; similarly, if at least $n-1$ members rank a candidate c_i higher than a candidate c_j, then it must be that c_i is more skilled than c_j. Hence, if the preference profile were known then a SCF can not only deduce the highest skilled candidate, but can also rank the candidates according to skill.

Nash equilibrium. There are always $n-1$ members who can each deviate from any Nash equilibrium under Rules 2 or 3 and enforce, via Rule 3, any outcome. Suppose R is the true preference profile and $(a,R^{'},z)$ is a NE under Rule 1. By implementation a is the highest skilled candidate under R and $R^{'}$ (if it is not the highest skilled under $R^{'}$ then it would not be in $F(R^{'})$ and $(a,R^{'},z)$ would not fall under Rule 1). Thus, the least skilled candidate is less preferred than a by at least $n-1$ members under $R^{'}$. Rule 2 allows any member to deviate and get any outcome that is less preferred for that member under the preferences reported by others. Hence, at least $n-1$ members can *each* deviate to Rule 2 and force the mechanism to select the least skilled candidate. Therefore, it takes only one member reporting the least skilled candidate for the mechanism to pick the least skilled candidate instead of the highest skilled!

In practice such a failure can arise for many reasons: a member may intentionally try to harm others; a member may misjudge the least skilled candidate to be the highest skilled; the least skilled candidate may be able to deceive one member into thinking they are highly skilled; or simply due to a mistake or a failure in communication in reporting the messages.

The lack of safety illustrated in this example is worrisome. In Maskin's mechanism and many others in the literature, it is generally true that a deviation by one agent, for any reason, can lead to unrestricted damage to others.

2.2 A General Notion of Safety

We first introduce some notation that will be used throughout the paper. Let $A = \{a_1,\ldots,a_x\}$ be the set of outcomes and $I = \{1,\ldots,n\}$ be the set of players. Each individual $i \in I$ has a preference relation R_i over the set of outcomes, A. The collection of preference relations for all individuals is called a preference profile and is denoted by $R \equiv (R_1,\ldots,R_n)$. Associated with each preference relation R_i is a strict preference relation P_i. The set of all permissible preference profiles is denoted by $\mathscr{R}$ and we let $\mathscr{R}_i$ be the set of all permissible preference relations for a specific player $i \in I$. A *mechanism* is a pair (M,g), where $M = \times_{i\in I} M_i$ is the product of message (or strategy) spaces for each player and $g : M \to A$ is an *outcome function*. A typical element of M_i is a message for player $i \in I$, denoted by m_i, and a collection of such messages is called a message profile and denoted by $m = (m_1,\ldots,m_n)$. For $B \subseteq I$, let $M_B := \times_{i\in B} M_i$ and let $(m_B^{'}, m_{-B})$ denote a message profile in M with players in B reporting $m_B^{'} \in M_B$ and those in $I\backslash B$ reporting $m_{-B} \in M_{I\backslash B}$. A mechanism (M,g), together with a preference profile R defines a normal form *game*. For $\alpha \in \{1,\ldots,n\}$, let I_α be the set of all subsets of I of size α.

Intuitively, our goal is to design mechanisms where players are safe. The American Heritage dictionary defines "safe" simply as "free from risk". Risk can be viewed as exposure to

harm. We must then define what it means for a player to be harmed. For generality, we do not make that definition explicit yet; we simply refer to a "harmed" player as a *victim*.

If safety is freedom from exposure to harm (i.e. freedom from risk), players must be limited in their ability to victimize others. It will thus be helpful to define a function that quantifies the ability of players to victimize others. We do this using the following definition of a *victims function*:

Definition 1. Let $m \in M$ be a strategy profile. The victims function at m is a function $V_m : \{1,\dots,n-1\} \to \{0,\dots,n-1\}$ defined by

$$V_m(\alpha) = \max_{B \in I_\alpha} \left[\max_{m'_B \in M_B} \left(\sum_{i \in I \setminus B} \mathbf{1}\{\text{i is a victim}\} \right) \right]$$

The victims function specifies the maximum number of victims that α players, which we refer to as *deviators*, can create by potentially deviating from the profile under consideration. We can now define safety in terms of the victims function.

Definition 2. Given a game, (M, g, R), a strategy profile, $m \in M$, is (α,ω)-safe if $V_m(\alpha) \leq \omega$.

A profile is (α,ω)-safe if no group, $B \subseteq I$, of α players can obtain an outcome (by potentially deviating from m) that leads to more than ω victims in $I \setminus B$ under R. Different combinations of α and ω can lead to different safety levels that may be of interest depending on the context. Intuitively, a profile is safe if players are limited in their ability to victimize others, leading to a formal definition of safety as freedom from exposure to harm.

Note that if $\omega \geq n - \alpha$ then the safety requirement imposes no restrictions. Throughout the rest of the paper, unless specified otherwise, ω will always assumed to be strictly less than $n - \alpha$.

Definition 2 is extremely general. Using variations on the definition of a victim, we can use definition 2 to examine the safety properties of virtually any game. Section 8.1 discusses this point more in depth. To make this definition applicable in particular contexts, we must define what it means to be a victim more precisely, which we do next.[3]

2.3 A Notion of Victimization

We adopt a notion of victimization that enables us to apply our safety concept to the standard implementation problem. In much of the mechanism design and implementation literature, a player mainly cares about how the final outcome of the mechanism ranks in that player's

[3]We note the inherent binary nature of the concept of a victim: a player is either a victim or not a victim. At first glance, it may seem like this limits the applicability of our safety definition in situations where a designer cares about the level of damage players are exposed to. However, this can be accommodated by letting the definition of a victim depend on the level of damage and examining the safety properties at different damage thresholds. We allow for this possibility in our definition of a victim.

preferences relative to other outcomes in some finite set.[4] Thus, we can say that a player is a victim if the final outcome of the mechanism ranks too low relative to other outcomes in that player's preferences. Formally, a player is a victim if the outcome of the mechanism is among the K least preferred outcomes for that player, where $K \in \{1,\dots,x-1\}$ is a threshold chosen by the designer. This leads to the following definition.

Definition 3. Given a preference profile R and an outcome $a \in A$, player i is a victim if

$$|\{b \in A : aR_i b\}| \leq K \tag{1}$$

If (1) holds we also say that a is among the K least preferred outcomes for player i under R. Using this definition of a victim, we can write the victims function as,

$$V_m(\alpha) = \max_{B \in I_\alpha} \left[\max_{m'_B \in M_B} \left(\sum_{i \in I \setminus B} \mathbf{1}\left\{ |\{b \in A : g(m'_B, m_{-B}) R_i b\}| \leq K \right\} \right) \right]$$

Definition 3 has subtle implications for cases where players are indifferent between some outcomes, as the following example illustrates.

Example 1. Suppose $A = \{a_1, a_2, a_3\}$, $K = 2$, and there are 3 players with the following preferences:

R_1	R_2	R_3
a_2	a_1, a_2	a_1, a_2, a_3
a_3	a_3	
a_1		

Then, the set of least preferred K outcomes for player 1 is $\{a_1, a_3\}$, for player 2 is $\{a_3\}$, and for player 3 is $\varnothing$. Hence, players who are indifferent between some outcomes will, in general, have smaller sets of outcomes that lead to their labeling as victims; players who are indifferent between all outcomes cannot be victimized.

It is easy to see why definition 2, using definition 3 of a victim, leads to desirable safety properties in an implementation framework. If a strategy profile is (α, ω)-safe, with victimization defined using definition 3, then no group of α players can deviate and cause more than ω of the rest to get an outcome that ranks too low (among the least K favorite) in their preferences. Thus, groups of deviators cannot significantly hurt others, ruling out "disasters" and limiting how much players can reduce the wellbeing of others. By choosing α, ω, and K, the designer has great flexibility and control over the desired level of safety. Section 8.1 discusses our notions of safety and victimization in more detail.

[4] We can easily accommodate infinite infinite sets of outcomes as we explain in section 8.1.

Note: throughout the paper we suppress the dependence on the damage threshold, K, to simplify the notation. However, K is used to define a victim (definition 3), so any discussion of safety implicitly assumes a given $K \in \{1,\ldots,x-1\}$.

3 Strong Safe Implementation

As before, let the set of all permissible preference profiles be denoted by $\mathscr{R}$ and a game be a mechanism (M,g) together with a preference profile $R \in \mathscr{R}$. A *social choice rule* (SCR) or a *social choice correspondence* (SCC) is a function $F : \mathscr{R} \to 2^A \setminus \{\varnothing\}$. When F is single valued, i.e. $F : \mathscr{R} \to A \setminus \{\varnothing\}$, it is called a *social choice function* (SCF). A *solution concept* is a prediction specifying the strategy profiles, called *equilibria*, we expect players to play in a game. If S is a solution concept and $\Gamma = (M,g)$ is a mechanism, we denote the set of all equilibrium profiles under preference profile R by $S^\Gamma(R) \subseteq M$. The set of outcomes associated with $S^\Gamma(R)$ is denoted by $g(S^\Gamma(R))$. A SCC (or SCF) is implementable in a solution concept, S, if there exists a mechanism $\Gamma = (M,g)$ such that $g(S^\Gamma(R)) = F(R)$ for all $R \in \mathscr{R}$. We use NE to denote the use of Nash equilibrium as a solution concept.

Throughout the rest of the paper, referring to a strategy profile as "safe" implies that it is (α,ω)-safe where $\alpha \in \{1,\ldots,n-1\}$ and $\omega \in \{0,\ldots,n-2\}$ are set by the designer.

We first consider a strong form of safe implementation where we require the implementing mechanism to have the following properties: all Nash equilibria lead to desirable outcomes, all desirable outcomes can be obtained via Nash equilibria, and all Nash equilibria are safe.

Definition 4 (Strong (α,ω)-Safe Implementation). A social choice rule is *strongly* (α,ω)-*safe implementable* if a mechanism $\Gamma = (M,g)$ exists such that for any preference profile $R \in \mathscr{R}$, $g(NE^\Gamma(R)) = F(R)$ and all profiles in $NE^\Gamma(R)$ are (α,ω)-safe.

Strong safe implementation is simply Nash implementation with the added constraint that all Nash equilibria of the mechanism at any preference profile are safe.

Strong safe implementation is desirable because it implies that players will be safe as long as they play a Nash equilibrium. Unfortunately, the following Theorem shows that strong safe implementation implies either very strong restrictions on the implementable social choice rule or strong and unnatural restrictions on preferences that are unlikely to hold except in very specific environments. The proof for the theorem is in the appendix.

Theorem 1. Let F be a social choice rule that is strongly (α,ω)-safe implementable and let $\mathscr{R}$ be the set of admissible preference profiles. Suppose $a \in F(R)$ for some $R \in \mathscr{R}$. Then at least one of the following conditions is true:

(1) $a \in F(R')$ for all $R' \in \mathscr{R}$.

(2) If $a \notin F(R')$ for some $R' \in \mathscr{R}$, then there exists a pair $(i,b) \in I \times A$ such that:

- $aR_i b$.

- $bP_i^{'}a$.
- There is no preference profile where a weakly rises in every player's preferences relative to R and b victimizes more than ω players in $I\backslash\{i\}$.

Condition (1) of Theorem 1 says that a, a desirable social outcome at some preference profile, must be a chosen at all other preference profiles; this is a very strong condition on the social choice rule. If condition (1) is true for all outcomes in the range of F then F is a trivial social choice rule: it is constant and does not depend on the preference profile.

Condition (2) says that if a is not picked at some other preference profile, $R^{'}$, then there must be an outcome that is weakly less preferred by some player i at R, strictly more preferred by player i at $R^{'}$, and never victimizes too many players at *any* preference profile where a rises in everyone's rankings; this must hold for every profile that does not contain a as an outcome. This is a strict and unnatural requirement on the preferences.

The most restrictive aspect of condition (2) is that relative preference movements of one outcome impose restrictions on *other* outcomes. For any outcome a and player i, why should the designer a priori believe that some outcome (that is sometimes weakly less preferred to a by player i and sometimes strictly more preferred) is never too low in the rankings of other players whenever a rises in everyone's rankings?

To see how restrictive conditions (1) and (2) are, we can consider some of their implications in the hiring example of section 2.1. We let $\alpha = \omega = K = 1$ so members are victims if their least preferred candidate is hired, and $\omega = 1$ guards against the least skilled candidate (who would victimize at least $n-1$ members) being selected. The preferences of the members in that example are simply determined by the relative skills of the candidates and the premiums members give for candidates in their fields. Thus, the conditions of Theorem 1 imply restrictions on the relative skills of the candidates that would presumably be known by the designer ex-ante:

(1) Some candidate, a, is known to always be the highest skilled candidate. If this is known ex-ante then there is no need for the hiring mechanism.

(2) Some candidate, a, may or may not be the highest skilled candidate, but if a is the highest skilled in more than one preference profile (so that, for example, relative skills of other candidates may vary), then at least one other candidate is ruled out from being the lowest skilled in profiles where a is the highest skilled.

There is no reason to think that those conditions would be known ex-ante by the designer. Thus, a designer cannot strongly safe implement a social choice function that picks the highest skilled candidate and at the same time allow for a variety of possibilities in relative candidate skills. This fact is not special to this example; it is much more general, as the following corollary shows.

Corollary 1. Suppose the set of preference profiles, $\mathscr{R}$, includes all possible strict preferences. Then no social choice rule is strongly safe implementable.

The assumption that $\mathscr{R}$ includes all possible strict preferences can be relaxed and the corollary would still hold. Intuitively, suppose $a \in F(R)$ for some $R \in \mathscr{R}$. If there is just one preference profile that violates condition (2) of Theorem 1 then a must be chosen at all preference profiles. If condition (1) of Theorem 1 holds and a victimizes more than ω players at *some* preference profile, strong safe implementation will fail because a is always chosen. Thus, the corollary holds as long as there are preferences where condition (2) of Theorem 1 is violated for some outcome, and as long as that outcome can potentially victimize many players.[5]

Note that corollary 1 holds for any arbitrary nontrivial combination of α, ω, and K, so it is true even with the weakest safety requirements.

4 Safe Implementation

Section 3 showed that strong safe implementation is restrictive and impossible in many environments. In this section we discuss a weaker notion of implementation, which we call (α,ω)-safe implementation. We provide necessary conditions for (α,ω)-safe implementation and show that they are also sufficient with mild preference restrictions when $\alpha < n-1$.

4.1 (α,ω)*-Safe Implementation*

For (α,ω)-safe implementation, we require the following two properties at any preference profile: all desirable outcomes can be obtained via Nash equilibria that are safe, and all Nash equilibria that are safe lead to desirable outcomes. This is equivalent to implementation in a solution concept that we refer to as a (α,ω)-safe equilibrium ((α,ω)-SE or simply SE):

Definition 5 (**(α,ω)-Safe Equilibrium**)**.** A (α,ω)-safe equilibrium is a strategy profile $m \in M$ that is both a Nash equilibrium and (α,ω)-safe.

Definition 6 (**(α,ω)-Safe Implementation**)**.** A social choice rule is (α,ω)-*safe implementable* if a mechanism $\Gamma = (M,g)$ exists such that for any preference profile $R \in \mathscr{R}$, $g(SE^{\Gamma}(R)) = F(R)$.

Safe implementation assures the designer that any outcome in the social choice rule can be obtained via a safe equilibrium of the mechanism. It is also logically consistent in the sense that if we expect players to play a Nash equilibrium that is also safe (i.e. a safe equilibrium),

[5] A less general version of corollary 1 for social choice functions follows directly from the Muller-Satterthwaite (1977) Theorem, which states that, under some conditions, Maskin monotonic social choice functions are dictatorial. Theorem 1 allows for more generality in corollary 1 and makes clear how the assumption that $\mathscr{R}$ includes all possible strict preferences can be relaxed.

then all safe equilibria should lead to outcomes in the social choice rule at any preference profile. Thus, it forms the basis for implementing social choice rules in a safe way.[6]

4.2 Necessary Conditions

This section shows that three conditions: ω-safe outcome, (α,ω)-safe monotonicity, and (α,ω,n)-similarity are necessary for implementation in (α,ω)-safe equilibria. The first two are necessary conditions on social choice rules; the third is a necessary condition on the set of preference profiles when $\alpha \geq \frac{n}{2}$. We start first with the ω-safe outcome property, which intuitively follows from the (α,ω)-SE definition.

Definition 7 (ω**-Safe Outcome Property).** If $a \in F(R)$ then at most ω players have a as one of the least K favorite outcomes under R.

It is easy to see why the ω-safe outcome property is necessary for implementation: if implementation holds and it is not satisfied then there must be a (α,ω)-SE of the implementing mechanism at which there are more than ω victims, directly contradicting the definition of (α,ω)-SE.

An implicit implication of the ω-safe outcome property is that the objectives of the designer (the social choice rule) are consistent with choosing an outcome that is not too low in the rankings of many players. This is not restrictive in most of mechanism design, which generally aims to combine individual preferences to reach a collective decision in some welfare maximizing way.

The ω-safe outcome property is similar to the "no worst alternative" property introduced by Cabrales and Serrano (2011). Cabrales and Serrano say that a SCC satisfies the no worst alternative property if, for all players, any outcome chosen in any preference profile is strictly preferred to some other outcome. If $\omega = 0$, $K = 1$, and each player has a single worst alternative then the ω-safe outcome property is equivalent to the no worst alternative property.

The next property, (α,ω)-safe monotonicity, is a generalization of Maskin monotonicity (Maskin (1999)). (α,ω)-Safe monotonicity changes depending on α with different general conditions for the case of $\alpha = 1$ than other cases. The intuition is that, at a (α,ω)-SE, unilateral deviations must satisfy incentive (Nash) and safety constraints, whereas multilateral deviations need to satisfy safety constraints only.

Definition 8 ((α,ω)**-Safe Monotonicity Property).** For any two profiles R, R' suppose $a \in F(R)$ and the following two conditions hold:

[6] In this paper, we do not simply assume that players coordinate on a safe equilibrium. However, to the extent that a safe equilibrium may be an appropriate prediction in some environments (e.g. as in Halpren (2008)'s t-immunity), implementation in safe equilibria may be interesting in its own right.

(1) For all $i \in I$, if $aR_i y$ and y victimizes at most ω players in the set $I \backslash \{i\}$ under R, then $aR_i^{'} y$ and y victimizes at most ω players in the set $I \backslash \{i\}$ under $R^{'}$.
(2) If $\alpha \geq 2$ then for all groups $B \subseteq I$ of size $n - \alpha$, if $x \in A$ victimizes at most ω players in B under R, then x victimizes at most ω players in B under $R^{'}$.

then $a \in F(R^{'})$

The first condition of the (α, ω)-safe monotonicity property states that for any player, i, if an element in A is weakly less preferred to a under R and causes at most ω victims in the set of all players excluding i, then the same holds true for that element under $R^{'}$. The second condition states that for any group B of size $n - \alpha$, if an element victimizes at most ω players in B under R then that element victimizes at most ω players in B under $R^{'}$.

The following Theorem shows that both the ω-safe outcome property and (α, ω)-safe monotonicity are necessary for implementation in (α, ω)-safe equilibria.

Theorem 2. Suppose F is implementable in (α, ω)-safe equilibrium. Then F satisfies the ω-safe outcome property and the (α, ω)-safe monotonicity property.

It is, again, easy to see why (α, ω)-safe monotonicity is necessary. Suppose F is (α, ω)-safe implementable and $a \in F(R)$ then there must be a (α, ω)-SE, call it m, at R with a as the outcome. If both conditions of (α, ω)-safe monotonicity are satisfied then m is a (α, ω)-SE at $R^{'}$ as well. Hence, $a \in F(R^{'})$. A formal proof is provided in the appendix.

To understand (α, ω)-safe monotonicity better, we can express it differently. For any two profiles R, $R^{'}$ suppose $a \in F(R)$ but $a \notin F(R^{'})$. Then F is (α, ω)-safe monotonic if one of the following conditions hold:

(1) For some $i \in I$ and $y \in A$, $aR_i y$ and y victimizes at most ω players in the set $I \backslash \{i\}$ under R, but either $yP_i^{'} a$ or y victimizes more than ω players in the set $I \backslash \{i\}$ under $R^{'}$.
(2) $\alpha \geq 2$ and there is some group $B \subseteq I$ of size $n - \alpha$ and some $x \in A$ such that x victimizes at most ω players in B under R but victimizes more than ω players in B under $R^{'}$.

It is useful to compare (α, ω)-safe monotonicity with Maskin monotonicity. Maskin monotonicity says that if an outcome is chosen at a particular profile and this outcome is not chosen in another profile then some player must have had a preference reversal where this outcome fell relative to another between the two profiles.

(α, ω)-Safe monotonicity is equivalent to Maskin monotonicity when $\alpha = 1$ and $\omega = n - \alpha$. In this case the second condition in (α, ω)-safe monotonicity never plays a role and it is impossible for a player to victimize more than ω of the rest so the first condition reduces to simple preference reversals, which is the essence of Maskin monotonicity. For more general combinations of α and ω, Maskin monotonicity does not imply (α, ω)-safe monotonicity and

(α,ω)-safe monotonicity does not imply Maskin monotonicity, but both have a nonempty intersection. The following examples illustrate these facts.

Example 2. Let $A=\{a,b,c\}$ and $K=\alpha=\omega=1$ and consider a social choice function given by $F(R)=a$ and $F(R')=c$, where R and R' are given by

R_1	R_2	R_3	R_1'	R_2'	R_3'
a	a	b	c	c	a,b,c
b	b	a	a	a	
c	c	c	b	b	

F satisfies Maskin monotonicity trivially because players 1 and 2 have preference reversals between the chosen outcome and a less preferred one across profiles in both directions.

Next, we check (α,ω)-safe monotonicity. Note that because $\alpha=1$, the second condition of (α,ω)-safe monotonicity does not play a role. Also, since $K=1$, a player is victimized by some outcome if that outcome is the least preferred for that player. Because $a\in F(R)$, condition (1) of (α,ω)-safe monotonicity holds (going from R to R') if for each player i, the set of outcomes weakly less preferred to a that victimize at most $\omega=1$ players in $I\backslash\{i\}$ weakly grows going from R to R'. We check this next:

- For player 1 at R, the set of outcomes weakly less preferred to a that victimize at most one of the players in the set $\{2,3\}$ is $\{a,b\}$. At R' that set is $\{a,b\}$.
- For player 2 at R, the set of outcomes weakly less preferred to a that victimize at most one of the players in the set $\{1,3\}$ is $\{a,b\}$. At R' that set is $\{a,b\}$.
- For player 3 at R, the set of outcomes weakly less preferred to a that victimize at most one of the players in the set $\{1,2\}$ is $\{a\}$. At R' that set is $\{a,c\}$.

Thus, condition (1) of (α,ω)-safe monotonicity holds going from profile R to R'. However, $a\notin F(R')$, so (α,ω)-safe monotonicity is not satisfied and F is not (1,1)-safe implementable with $K=1$. Hence, Maskin monotonicity does not imply (α,ω)-safe monotonicity.

Note: if $F(R')$ is modified to be $\{a,c\}$, then (α,ω)-safe monotonicity is satisfied, showing that Maskin monotonicity and (α,ω)-safe monotonicity have a nonempty intersection.

Example 3. Let $A=\{a,b,c\}$ and $K=\alpha=\omega=1$ and consider a social choice function given by $F(R)=a$ and $F(R')=c$, where R and R' are given by

R_1	R_2	R_3	R_1'	R_2'	R_3'
a	a	b	a,c	a,c	a,c
b	b	a	b	b	b
c	c	c			

Maskin monotonicity is not satisfied: a weakly rises in all players' ranking going from R to R', but $a \notin F(R')$. However, (α,ω)-safe monotonicity is trivially satisfied: for player 1 at R, the set of outcomes weakly less preferred to a that victimize at most one of the players in the set $\{2,3\}$ is $\{a,b\}$. At R' that set is $\{a,c\}$. Because $\{a,b\} \not\subseteq \{a,c\}$, condition (1) of (α,ω)-safe monotonicity fails going from R to R'. Going from R' to R, a is weakly less preferred to c and victimizes no players at R', but it is strictly more preferred to c by all players at R. Thus, condition (1) also fails going from R' to R and (α,ω)-safe monotonicity holds trivially. This shows that (α,ω)-safe monotonicity does not imply Maskin monotonicity.

Finally, we define a notion of similarity among the elements in a set of preference profiles that will be necessary for safe implementation only when $\alpha \geq \frac{n}{2}$.

Definition 9 (**(α,ω,n)-Similarity).** Given α, ω, and n, let $q = \left\lfloor \frac{n}{n-\alpha} \right\rfloor$ be the integer quotient from dividing n by $n-\alpha$ (and ignoring the remainder). A set of preference profiles $\mathscr{R}$ is (α,ω,n)-similar if whenever $\{(g^1,R^1),\ldots,(g^q,R^q)\}$, where $g^i \subseteq I$ and $R^i \in \mathscr{R}$, are any q pairs such that $|g^i| \geq n-\alpha$ and $g^i \cap g^j = \varnothing$ for any $i,j \in \{1,\ldots,q\}$, then there is at least one outcome in A that causes ω or less players in g^i to be victims under R^i regardless of $i \in \{1,\ldots,q\}$.

If a set of preference profiles, $\mathscr{R}$, is (α,ω,n)-similar then the preference profiles in $\mathscr{R}$ are not too different from each other. Specifically, if the players are partitioned in a way to maximize the number of groups of size $n-\alpha$ and each of those groups reported a different preference profile from a (α,ω,n)-similar set, then we can find an element that victimizes at most ω players in each group under the profile reported by that group. This condition is necessary when $\alpha \geq \frac{n}{2}$ because for (α,ω)-safe implementation, any group of size $n-\alpha$ or larger must be able to protect its members, except at most ω of them, from being victims. If $\alpha \geq \frac{n}{2}$ then the deviators can act as if they were a group of $n-\alpha$ or more "rational" players. The designer in this case would have no way of knowing which group of players to protect, so the designer must make sure that all groups of size $n-\alpha$ or greater are protected (in the (α,ω) sense). If the set of preference profiles is (α,ω,n)-similar then the designer can assure the safety of any group of size at least $n-\alpha$.

The (α,ω,n)-similarity is a somewhat restrictive constraint on preferences, but the following Theorem establishes its necessity when $\alpha \geq \frac{n}{2}$; it is not necessary for safe implementation when $\alpha < \frac{n}{2}$. The proof for the Theorem is in the appendix.

Theorem 3. Suppose F is implementable in (α,ω)-safe equilibrium with n players on a set of preference profiles $\mathscr{R}$ and suppose that $\alpha \geq \frac{n}{2}$. Then $\mathscr{R}$ is (α,ω,n)-similar.

4.3 Sufficient Conditions

In this section we provide preference restrictions that, along with the necessary conditions presented in the previous section, are sufficient for implementation in (α,ω)-safe equilibria when $\alpha < n-1$.

Definition 10 (**(α,ω)-Exposure**). A preference profile $R \in \mathscr{R}$ satisfies (α,ω)-exposure if for any subset of players $B \subseteq I$ of size $n-\alpha$, the set of outcomes in A that cause more than ω players in B to be victims is nonempty.

(α,ω)-Exposure rules out profiles where some groups of α players cannot victimize more than ω of the other players even if the α deviators were allowed to pick any outcome in A. These are profiles where the question of safety is vacuous for some groups: profiles where some groups of $n-\alpha$ or more players are "invincible" in the sense that no outcome in A can victimize more than ω in those groups (e.g. profiles where all players are indifferent between all outcomes). In economic environments, where any group of α deviators can choose to deprive all the rest from all economic goods and split the goods among themselves, the (α,ω)-exposure condition is not restrictive.

The following Theorem is the main result of the paper; it establishes that the ω-safe outcome and the (α,ω)-safe monotonicity properties are necessary and sufficient for (α,ω)-safe implementation under the appropriate preference restrictions that depend on α.

Theorem 4. Suppose F satisfies the ω-safe outcome property and the (α,ω)-safe monotonicity property, admissible preference profiles satisfy (α,ω)-exposure, and $n > 2$. If $\alpha < \frac{n}{2}$, or if $\frac{n}{2} \leq \alpha < n-1$ and the set of preference profiles is (α,ω,n)-similar, then F is (α,ω)-safe implementable.

The proof for Theorem 4 is constructive and is provided in detail in the appendix. Here, we describe the mechanism used for implementation and outline the steps of the proof. The mechanism depends on whether $\alpha < \frac{n}{2}$ or $\frac{n}{2} \leq \alpha < n-1$. We describe first the mechanism for the case of $\alpha < \frac{n}{2}$. Each player reports an outcome in A, a preference profile from $\mathscr{R}$, and an integer in $\{1,\ldots,n\}$. Hence, the strategy space for each player is $A \times \mathscr{R} \times \{1,\ldots,n\}$, and a message for player i is denoted by (a^i,R^i,z^i). The outcome function, $g(\cdot)$, is determined as a function of the strategy profile, m, as follows:

Rule 1: If m is such that all players agree on a report, say (a,R,z), and if $a \in F(R)$ then $g(m) = a$.

Rule 2: Suppose all players agree on (a,R,z) with $a \in F(R)$, except for one player i who reports $(a^i,R^i,z^i) \neq (a,R,z)$. If a^i is weakly less preferred than a for player i under R and if a^i does not victimize more than ω players in the set $I\backslash\{i\}$ under R then $g(m) = a^i$; otherwise, $g(m) = a$.

Rule 3: Suppose $\alpha \geq 2$ and all players agree on (a,R,z), except for a set of players B who submit different reports and $1 < |B| \leq \alpha$. Let B' be the set of players in B who report an outcome that does not victimize more than ω players in $I \backslash B$. If B' is not empty, then sort the players in B' from lowest to highest according their index, rename player $i \in B'$ by the new sorted order $q_i \in \{1,\ldots,|B'|\}$, and let $g(m) = a^{q_i^*}$, where $q_i^* = 1 + ((\sum_{i \in B'} z^i) \mod |B'|)$ is the q_i^*'th player in B'; otherwise $g(m) = a$.

Rule 4: In all other cases, $g(m) = a^j$ for $j = 1 + ((\sum_{i \in I} z^i) \mod n)$.

Rule 1 allows players to get any outcome in F if they all agree on the report submitted. Rule 2 allows a player to deviate from a consensus and receive any outcome if that outcome is weakly less preferred for that player and if that outcome does not victimize more than ω of the others under the profile reported by the other players. Rule 3 allows a set of players of size at most α to deviate and receive any outcome as long as that outcome does not victimize more than ω of the other players under the profile reported by the others. Finally, Rule 4 is a modulo game allowing any player to get any outcome they desire.

The proof must show that any outcome in F can be obtained as the outcome from a (α,ω)-SE of the mechanism above and any (α,ω)-SE of the mechanism maps into F. The first part is simple; suppose R is the true preference profile and suppose $a \in F(R)$. The profile m given by $m_i = (a,R,1)$ is a (α,ω)-SE of the mechanism under Rule 1 that leads to $g(m) = a$. To see that m is a (α,ω)-SE, note that no player has an incentive to deviate due to Rule 2, and no set of players of size at most α can victimize more than ω of the rest due to Rules 2 and 3. Hence, any outcome in F can be obtained as the outcome from a (α,ω)-SE of the mechanism.

An observation crucial to completing the second part of the proof is that no (α,ω)-SE of the mechanism above can exist under Rules 2, 3, or 4. This is because under these rules there is always a set of players of size α that can, via Rule 4, obtain any outcome in A. By assumption, the preference profile satisfies (α,ω)-exposure so there is some outcome in A that victimizes more than ω players in each set of players of size $n - \alpha$. This implies that a set of α players can victimize more than ω of the rest, violating the definition of a (α,ω)-SE. Hence, the only (α,ω)-safe equilibria of the mechanism are under Rule 1. Now, for any (α,ω)-SE $(m_i = (a',R',z'))$ under Rule 1, Rules 2 and 3 imply that the conditions of (α,ω)-safe monotonicity are satisfied for R, which shows that $a' \in F(R)$.

If $\frac{n}{2} \leq \alpha < n-1$ the proof remains unchanged except that the set of possible profiles is assumed to be (α,ω,n)-similar and rule 3 of the mechanism is modified to be:

Rule 3: A group of players is called a *consensus group* if all the players in that group agree on a report and no player outside the group agrees with the players in the group. Suppose rule 2 does not apply and suppose there are *p consensus groups*, each of size $n - \alpha$ or more players.

- Case 1. If there is only one consensus group, call it B_1, then let $g(m) = a^q$ for $q = \max\{j \in I \backslash B_1\}$ if no element in the set $\{a^j : j \in I \backslash B_1\}$ victimizes more than ω players in set B_1 under the profile reported by players in B_1; otherwise set $g(m) = \bar{a}^1$, where $\bar{a}^1$ is the outcome reported by players in B_1.
- Case 2. If there are multiple consensus groups then pick an outcome that does not victimize more than ω players in any of the consensus groups under the profiles reported by the players in the groups. Such an outcome exists because of the assumption that the set of preference profiles is (α, ω, n)-similar.

The function of the first case in the modified rule 3 is to allow a group of α deviators from rule 1 to achieve any outcome they desire as long as that outcome does not victimize more than ω players of the rest. This is needed to allow the (α, ω)-safe monotonicity condition to work. The second case is for profiles where it is impossible for the designer to distinguish between the deviators and the non-deviators. We use the assumption that the set of preference profiles is (α, ω, n)-similar to allow the designer to pick an outcome that does not victimize more than ω players in any consensus group of size $n - \alpha$ or more.

Using a modulo game in the proof is very helpful in establishing a theorem with the generality of Theorem 4; it allows us to focus on the necessary and sufficient conditions for (α, ω)-safe implementation rather than on restrictions that arise due to requirements on the mechanisms used in implementation. Integer, or modulo, games are often used in the implementation literature to rule out undesirable equilibria. In general problems it is difficult to rule out undesirable equilibria without allowing players a systematic way of deviating and obtaining more preferred outcomes; in smaller problems or more specific examples this may be possible without using a complex mechanism, but a modulo game is helpful when considering an arbitrary social choice rule on an arbitrary set of preferences.

5 Double Implementation

By itself, safe implementation does not guarantee that all Nash equilibria will lead to desirable social outcomes, because it places no restrictions on unsafe Nash equilibria. In this section we explore double implementation in both safe and Nash equilibria.

First, we introduce a necessary monotonicity condition on social choice rules that must hold for double implementation when transfers are not allowed.

Definition 11 (ω**-Safe-Nash Monotonicity**)**.** For any two profiles R and R', suppose $a \in F(R)$ and the following condition holds: for all $i \in I$, if $aR_i b$ and b victimizes at most ω players in the set $I \backslash i$ under R then $aR'_i b$. Then $a \in F(R')$.

Theorem 5. Suppose F is double implementable in (α, ω)-safe and Nash equilibria. Then F is ω-safe-Nash monotonic.

ω-Safe-Nash monotonicity is more restrictive than Maskin and (α,ω)-safe monotonicity, and it implies both.[7] However, in the next section we show that if side payments are allowed and the ω-safe outcome property holds, then ω-safe-Nash monotonicity is sufficient for double implementation in safe and Nash equilibria when $\alpha < n/2$.[8]

5.1 Double Implementation with Transfers

In the rest of this section we allow the designer to specify a vector of transfers $t(m) = (t_1,\ldots,t_n) \in \mathbb{R}^n$ for each $m \in M$ in mechanism (M,g). We use the convention that transfers are paid to the players, so if t_i is negative then player i pays $|t_i|$. We impose no assumptions except that there exists a large enough transfer $L \in \mathbb{R}$ so that any player would prefer to avoid paying L (having transfer $-L$ as opposed to 0), and would prefer a transfer of L instead of 0, regardless of the outcome or the preference profile.

For the following, we extend preferences to be defined over sets of outcomes and transfers $A \times \mathbb{R}$. If $(a,0)R_i(b,0)$ then player i weakly ranks outcome a above b at preference profile R; we often drop the transfers if they are 0 to simplify notation. We assume positive transfers are valued positively by players at all preference profiles.

Before we proceed, we discuss some slight modifications that are needed in environments with transfers. First, to avoid trivially eliminating victims by compensating them for bad outcomes, we constrain our mechanisms throughout to have no transfers in equilibrium. Second, our definition of a victim requires a slight modification, because the notion of the least K preferred outcomes may not be well defined when negative transfers are allowed. We define a victim to be a player who obtains an outcome within their least K preferred outcomes among the outcomes in A without consideration for transfers or who has a negative transfer (or both). This definition also accounts for players attempting to cause harm by imposing negative transfers on others.

Theorem 6. Suppose $n > 2$, $\alpha < \frac{n}{2}$, and F satisfies ω-safe-Nash monotonicity and the safe outcome property. Then F is double implementable in (α,ω)-safe equilibria and Nash equilibria using a mechanism with transfers.

The proof uses the following transfer scheme modification to the mechanism used for safe implementation:

[7] To see this, we can write ω-safe-Nash monotonicity as follows: if $a \in F(R)$ but $a \notin F(R')$, then for some $i \in I$ and some $b \in A$, aR_ib and b victimizes at most ω players in the set $I\backslash i$ under R, but bP'_ia. This directly implies the contrapositive of (α,ω)-safe monotonicity (stated after Theorem 2), showing that (α,ω)-safe monotonicity holds. It also implies that if $a \in F(R)$ but $a \notin F(R')$ then at least one player had a preference reversal between a and a weakly lower ranked outcome at R going from R to R', showing that Maskin monotonicity holds.

[8] ω-safe-Nash monotonicity is, in general, not necessary for double implementation when mechanisms utilizing transfers are allowed, unless the designer does not have precise cardinal information, as we explain section 5.2.

Transfer Scheme 1

Rule 1: All transfers are zero under rule 1.

Rule 2: Suppose all players agree on (a,R,z) with $a \in F(R)$, except for one player i who reports $(a',R',z') \neq (a,R,z)$. If $R' \neq R$ and $z' \neq z$ then all transfers are zero; otherwise player i is fined L.

Rule 3: Under this rule $\alpha \geq 2$ and all players agree on (a,R,z), except a set of players B who submit different reports and $1 < |B| \leq \alpha$. Impose a fine L on all players in B unless the following is true: $|B| = 2$, call them players i and j, with player i reporting (a',R',z') and player j reporting (a'',R'',z''), with $R' \neq R$, $z' \neq z$, $R'' = R$, and $z'' = z'$, then impose a transfer L from player i *to* player j.

Rule 4: Each player reporting some $z \in \{1,\ldots,n\}$ obtains a transfer L from each player reporting $z-1$.

The proof of Theorem 6 shows that no Nash equilibria exist under rules 2,3, or 4. Under rule 2, player i can choose whether to be fined or not without affecting the outcome. If a Nash equilibrium exists under rule 2 then no players will be fined, but in that case the transfer scheme allows a "whistleblower" to reveal the unsafe situation by deviating to rule 3 and obtaining a transfer of size L from player i. Thus, no Nash equilibria exist under rule 2. The proof also shows that no Nash equilibria exist under rules 3 or 4.

The following fact shows that double implementation can actually bring us very close to strong safe implementation. In fact, it can bring us close enough that mild assumptions on behavior (unrelated to safety) can achieve strong safe implementation.

Corollary 2. Suppose F is double implementable in Nash and safe equilibria using the mechanism described in the proof of Theorem 6. Then the following is true:

(1) Any unsafe Nash equilibrium is untruthful (all players report a false preference profile).

(2) For any unsafe Nash equilibrium there exists a truthful, safe equilibrium that yields the same outcome as the unsafe Nash equilibrium.

Thus, players gain nothing by playing an unsafe Nash equilibrium over a safe one, and they all must lie to play an unsafe Nash equilibrium. Mild assumptions on collective honesty of players, or assumptions on truthful equilibria being focal when untruthful equilibria are payoff-equivalent are enough to achieve strong safe implementation, where players are always safe in equilibrium.

5.2 Ordinal Preference Information and Necessity of ω-Safe-Nash Monotonicity

We have shown that ω-safe-Nash monotonicity is necessary for double implementation in the absence of transfers. In this section we show that if the designer does not have access to

precise cardinal preference information then ω-safe-Nash monotonicity remains necessary, even if mechanisms with transfers are allowed. This will lead to necessary and sufficient conditions for double implementation.

Definition 12. Suppose $(a,0)R_i(b,0)$ and $(a,0)R'_i(b,0)$. A designer has *precise cardinal information* about preferences for player i and outcomes $\{a,b\}$ at preference profiles $\{R,R'\}$ if transfers $t_1,t_2 \in \mathbb{R}$ can be determined so that $(a,t_1)R_i(b,t_2)$ and $(b,t_2)P'_i(a,t_1)$, or if it can be determined that such transfers are do not exist.

Access to precise cardinal information implies a designer can determine if it is possible to set transfers that maintain a player's ranking of two outcomes at one preference profile, but reverse that ranking at another preference profile. Intuitively, this implies the designer knows how much more a player prefers one outcome to another at each of the two preference profiles.

Theorem 7. Suppose F does not satisfy ω-safe-Nash monotonicity, but F is double implementable in (α,ω)-safe and Nash equilibria. Then the designer has precise cardinal information about preferences for some $(i,a,b,R,R') \in I \times A \times A \times \mathscr{R} \times \mathscr{R}$, where $a \neq b$ and $R \neq R'$.

Theorem 7 implies that ω-safe-Nash monotonicity is necessary for double implementation if the designer does not have precise cardinal information about preferences. Thus, the results in this section establish that, when precise cardinal preference information is not available, ω-safe-Nash monotonicity and the ω-safe outcome property are both necessary *and* sufficient for double implementation when $n > 2$, $\alpha < n/2$, and mechanisms with transfers are allowed.[9]

6 No Guarantees and Safety Trade-offs

6.1 No Guarantees

An interesting case of safe implementation not covered in Theorems 4 or 6 is that of $\alpha = n-1$, and $\omega = 0$. If a social choice function is implementable in a $(n-1,0)$-SE then the designer can *guarantee* each agent an outcome other than the agent's least favorite K ones. Many institutions and laws attempt to guarantee players a minimum level of payoff if they follow particular strategies, regardless of what other players do. This section illustrates the difficulty with $(n-1,0)$-safe implementation in a wide range of environments when transfers cannot be used.

The following lemma provides a necessary condition on mechanisms that implement social choice functions in a $(n-1,0)$-SE.

[9]The necessity of the ω-safe outcome property follows directly from the definitions, regardless of whether transfers are allowed or not, given our constraint that no transfers exist in equilibrium and our notion of victimization.

Lemma 1. Suppose the set of preference profiles contains all strict preferences over A. Any mechanism, (M,g), that implements a social choice function in a $(n-1,0)$-safe equilibrium must be such that $\forall i \in I$ and for all K-size subsets $D \subseteq A$, $\exists m_{Di} \in M_i$ such that $g(m_{Di}, m_{-i}) \notin D$, $\forall m_{-i} \in M_{-i}$. [10]

The lemma states that any mechanism used in $(n-1,0)$-safe implementation must give every player possible strategies that allow him to rule out any K outcomes in A from being implemented regardless of the strategies of the other players. The proof of the lemma is in the appendix, but the intuition is simple. If a mechanism does not allow some player to rule out some K elements in A, then there will be a profile of preferences where those outcomes are the least preferred for that player but the other $n-1$ players have strategies that force the mechanism to select one of those outcomes, violating the conditions of $(n-1,0)$-SE.

The following theorem uses the previous lemma to establish an impossibility result for implementation of social choice functions in $(n-1,0)$-SE when preferences include all strict preference profiles.

Theorem 8. Suppose $n \geq 2$ and the set of preference profiles, $\mathscr{R}$, contains all strict preferences over A. If $x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$ then no social choice function is implementable in a $(n-1,0)$-safe equilibrium.

Note that the number of outcomes can be more than nK so the theorem does not rely on the fact that ruling out the least favorite K outcomes for each player could exhaust all the possibilities. Remarkably, this is true *regardless* of $K \in \{1,\ldots,x-1\}$; the theorem does not count on K being large enough. Thus, the planner cannot make guarantees even in cases where $K = 1$, the weakest guarantee possible.

What is most interesting about Theorem 8 is that it holds not precisely because a designer cannot make safety guarantees, but because of an interplay between guarantees and freedom that causes non-existence of Nash equilibria. Indeed, a designer can design a mechanism that keeps players safe even if $x > nK$: simply allow them strategies that rule out their least favorite K outcomes and pick an outcome that is not ruled out. However, the theorem shows that if $x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$ then such a mechanism will, in general, have no Nash equilibria because the power given to any player (to rule out any K outcomes) is too great that players may be able to use it to deviate and obtain better outcomes. The following example makes the intuition clear.

Example 4. Suppose a committee, $I = \{\text{Member 1, Member 2, Member 3, Member 4}\}$, of $n = 4$ members is deciding on hiring a candidate from among $x = 6$ candidates: $A = \{a_1,\ldots,a_6\}$. Let $K = 1$. In this case $x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$ and we know by Theorem 8 that no social choice function is implementable in a $(n-1,0)$-SE. Even though there are more candidates than

[10] This lemma can be easily extended to groups of players in general (α,ω)-safe implementation.

members, *no* mechanism can *guarantee* each member a candidate better than his least favorite. To see this clearly, let the preference profile, R, be as follows:

Member 1	Member 2	Member 3	Member 4
a_2	a_1	a_1	a_1
a_3	a_3	a_2	a_2
a_4	a_4	a_4	a_3
a_6	a_6	a_5	a_5
a_1	a_2	a_3	a_4
a_5	a_5	a_6	a_6

The choices under the solid line are those that victimize the members. Any potentially implementable social choice function in a $(n-1,0)$-SE must never pick a_5 or a_6 under R. So, WLOG assume that $F(R) = a_1$. To see that no mechanism can have a $(n-1,0)$-SE in which the outcome is a_1, let the message profile $m = (m_1, m_2, m_3, m_4)$ be a $(n-1,0)$-SE candidate. m_2 must rule out a_5 otherwise we could find m'_{-2} such that a_5 is forced as the outcome by three members at (m'_1, m_2, m'_3, m'_4) and m would not be a $(n-1,0)$-SE. Since m_2 ensures that a_5 will not be selected, Member 1 would rather deviate and choose a strategy that ensures that a_1 is not selected in order to get one of a_2, a_3, a_4, or a_6. We know that such a strategy exists by Lemma 1, so (m_1, m_2, m_3, m_4) is not an equilibrium because Member 1 has a profitable deviation.[11]

Despite the proof of Theorem 8 relying on richness of the preference space, the problem with non-existence of Nash equilibria does not disappear in environments with more restricted preferences. The essence of the problem is that each player will have "free" veto power over any $l \leq K$ outcomes if l of that player's least favorite outcomes are also among the K least favorite of some others. Thus, if a player does not like the outcome chosen, they will (under very weak preference conditions) be able to veto it. Even if $K = 1$, the weakest safety guarantee possible, weak preference restrictions will imply that in general many players can veto outcomes they don't like. This can easily become problematic when attempting to implement desirable social choice functions.

Note also that the assumption of all strict preferences being admissible can be relaxed. Lemma 1 and Theorem 8 only need that the planner does not ex-ante know which outcomes are among the least favorite $2K$ for each player. This can be seen from example 4; the first 4 rows of the preferences can be arbitrarily modified and the example would still hold.

[11] If Member 1 has a strategy that rules out more than one choice then a_1 must be ruled out because it is the second least favorite. Hence, $F(R)$ cannot be a_1, but $F(R)$ must be either a_1, a_2, a_3, or a_4, so we can assume that $F(R) = a_1$ without loss of generality.

Lemma 1 says that each player must have the freedom to rule out any K outcomes at a $(n-1,0)$-safe equilibrium. This can be seen as *negative freedom*: freedom from harm. Another kind of freedom we can explore is *positive freedom*: freedom of players to obtain different outcomes in equilibrium. In contrast to Lemma 1 which showed that safety requires the players to have certain negative freedoms, this section shows that safety restricts players' positive freedom. In general, *every* player will have no more positive freedom in a safe mechanism than in a mechanism that disregards safety. In addition, positive freedom becomes more restricted as mechanisms become safer.

Let the positive freedom of a player be parametrized by the number of options available to that player in equilibrium (i.e. the outcome selected at equilibrium and all outcomes the player can obtain by deviating). A player with more options in equilibrium has more positive freedom. Let $O_i(m)$ be the set of options available to some player i at message profile m and let $L_i(a,R) \equiv \{b \in A : aR_ib\}$ be the lower contour set of player i at outcome a when the preference profile is R (i.e. it is the set of all outcomes that are not strictly preferred by player i to a under R_i). Suppose the preference profile is R and a is chosen by the social choice rule at R. At any Nash equilibrium m that implements a of any mechanism, it must be true that $O_i(m) \subseteq L_i(a,R_i)$; that is, the set of options available to any player does not contain any outcome more preferred than the one selected in equilibrium. Any safe equilibrium, m', is a Nash equilibrium and has the property that no player can victimize more than ω players of the rest by deviating. Thus, letting $G_i(R)$ be the set of outcomes that do not victimize more than ω players in the set $I \setminus i$ under R, it must be true that $O_i(m') \subseteq L_i(a,R_i) \cap G_i(R)$. This shows that if mechanisms are designed to maximize the positive freedom of players subject to the desired constraints regarding incentives and/or safety then $|O_i(m')| \leq |O_i(m)|$ and a player has no more positive freedom at a safe equilibrium than at a Nash equilibrium.

Now suppose we generalize the definition of positive freedom to groups of players. Let $O_B(m)$ be the set of options available to a group of players B through a multilateral deviation from m (including the outcome when m is played). Nash implementation imposes no restrictions on $O_B(m)$, but at a safe equilibrium players in B must not be able to victimize more than ω of the rest if $|B| \leq \alpha$. Hence, as α increases, mechanisms that (α,ω)-safe implement social choice rules will be more restrictive in terms of the sets of options available to groups of players. Furthermore, as ω decreases and mechanisms become more safe, guaranteeing the victimization of less players, the set of outcomes that must be ruled out through deviations from equilibrium increases. This restricts the positive freedoms of all the players.

Using the notation in this section, the Appendix illustrates the connection between (α,ω)-safe monotonicity and the set of options available to a player i at equilibrium m, $O_i(m)$.

6.3 Safety vs. Efficiency

In some situations, the designer will face a choice between safe implementation and Nash implementation, and both will not be possible simultaneously. The advantage of safe implementation is that it guarantees that some Nash equilibria will be safe, and, more importantly, that every desirable outcome can be obtained by a Nash equilibrium. The drawback is that it imposes no restrictions on unsafe Nash equilibria, meaning that they may lead to undesirable outcomes outside of the social choice rule. Social choice rules are often chosen in a way to maximize overall welfare, so, to the extent that this is true, outcomes outside the social choice rule will be inefficient.

To illustrate, in section 7 we show that double implementation in safe and Nash equilibria is indeed possible if the designer has some "global" information about candidates' relative skill, but lacks "local" information. This is not a very restrictive assumption, but if the designer does not have this information, double implementation is impossible and the designer must choose between safe and Nash implementation. Nash implementation lacks the desirable properties of safe implementation, but has the advantage that every Nash equilibrium leads to the highest skilled candidate being selected. On the other hand, safe implementation allows for Nash equilibria that may select a less skilled candidate than the highest skilled. This trade-off between safety and efficiency extends more generally in situations where the designer must choose between safe and Nash implementation.

7 Example: Hiring with Multi-Dimensional Quality

For the motivating example in section 2.1, consider a mechanism that implements F in a (α,ω)-safe equilibrium with $\alpha = \omega = K = 1$. With the environment described for the problem, the least skilled candidate will be the least preferred for at least $n-1$ members. Hence, protecting members from deviations by any one ($\alpha = 1$) that cause them to get their least favorite outcome ($K = 1$) and allowing at most 1 member to be victimized ($\omega = 1$) results in protection from the least skilled candidate being selected through a one-member deviation.

We first show that the assumptions of Theorem 4 are satisfied.

Suppose $F(R) = a$ so that a is the candidate with the highest skill. Because there are at least 4 candidates in 4 different backgrounds and members rank candidates with backgrounds other than their own based on skill only, no member will have a as their least favorite outcome. Hence, the ω-safe outcome property is satisfied.

To show that (α,ω)-safe monotonicity holds, assume $F(R) = a$, the first condition of (α,ω)-safe monotonicity holds[12] moving from R to R', but (to show a contradiction) suppose $F(R') =$

[12] Only the first condition applies because $\alpha = 1$.

$b \neq a$. Because $\omega = 1$, the first condition of (α,ω)-safe monotonicity implies that any candidate that is not the least skilled at R and is weakly less preferred than a by some member must remain weakly less preferred than a by that member at $R^{'}$ and also not be the least skilled at $R^{'}$. Two implications follow. First, a is not the least skilled at $R^{'}$. Second, it must be that b was the least skilled at R, otherwise it could have only fallen relative to a in ranking at $R^{'}$ and a would be ranked higher than b by at least $n-1$ members at $R^{'}$ causing b to not be chosen at $R^{'}$. Because b is chosen at $R^{'}$, it must be more preferred than a by at least $n-1$ members. At $R^{'}$, the least skilled candidate is a candidate other than b or a, call it c. As mentioned earlier, the least skilled candidate will be the least preferred by at least $n-1$ members. Because a was the highest skilled at R, there must exist a member to whom c was less preferred than a at R and not the least preferred (because b is the lowest skilled candidate at R), but c becomes the least preferred at $R^{'}$, victimizing at least $n-1$ members. This violates the monotonicity condition leading to a contradiction.

Finally, we show that any preference profile arising from the environment described will satisfy (α,ω)-exposure. For any subset of members B of size $n-1$, the lowest skilled candidate must be ranked lowest by at least $n-2$ members in B. Because $n \geq 4$, the lowest skilled candidate victimizes at least 2 members in any group of size $n-1$, satisfying (α,ω)-exposure.

Because all the conditions of Theorem 4 are satisfied, we can directly use the mechanism given in the proof to (α,ω)-safe implement F. (α,ω)-safe implementation of F does not rule out the existence of undesirable Nash equilibria, but it implies that the highest skilled candidate can always be selected as an outcome from a safe profile, where deviations due to bad intentions, mistakes, or errors of judgment cannot cause disasters like selecting the lowest skilled candidate.

In this example, ω-safe-Nash monotonicity fails to hold, so double implementation is impossible[13] (this is true even if transfers are allowed, because we do not have precise cardinal preference information). However, suppose we restrict the preferences as follows: if $R, R' \in \mathscr{R}$ and b is the lowest skilled at R then b is not the highest skilled at R'. This condition implies that the designer has some information about relative skills of the candidates; the designer still does not know the relative skills "locally", but the designer has enough "global" information to rule out a candidate being the highest skilled if that candidate being the lowest

[13]To see this, consider two profiles R and R' with the property that candidate a is highest skilled at R, candidate b is the least skilled at R, and at R' all candidate skills remain the same, except that candidate b's skill becomes higher than all other candidates. Thus, $F(R) = a$ and $F(R') = b$. However, for any player i, any candidate that is less preferred than a at R and victimizes at most 1 player in $I \backslash \{i\}$ is less preferred than a for player i at R' because candidate skills do not change except for candidate b that victimizes at least 3 players. Thus, ω-safe-Nash monotonicity fails, because $a \notin F(R')$.

skilled is a possibility. If we allow such a restriction on preferences then ω-safe-Nash monotonicity is satisfied[14], and double implementation is possible using the mechanism in the proof of Theorem 6.

Double implementation in safe and Nash equilibria implies that safe implementation holds, and no unsafe Nash equilibria result in undesirable candidates. If, in addition, we are willing to assume that truthful equilibria, that are payoff-equivalent to untruthful ones, are focal then players will *always* be safe in equilibrium and the mechanism in the proof of Theorem 6 achieves strong safe implementation.

8 Discussion

Safety can be incorporated in mechanism design in various ways. This section discusses some important aspects of our safety concept and our notion of implementation.

8.1 Discussion of the Safety Concept

Our notion of safety is intuitive, it simply imposes restrictions on how much players can harm others. To make our definition more realistic and flexible, we allow for groups of players to attempt to cause harm and for some players to be victimized. In our notion of victimization, we also allow for different thresholds of safety by letting the designer choose the relative position in a player's ranking that defines a victim.

Our safety concept, as stated in definition 2, is very general. It applies to virtually any implementation or mechanism design problem. To illustrate this point in an auction setting, suppose we define a victim as *"a player with the highest private value who pays some amount of money but loses an auction due to irrational behavior by other players"*. Using this definition, it is easy to see that an auction with a reserve price is "safer" than one with an entry fee. Even though both auctions may be theoretically equivalent in some cases, an irrational player in an auction with an entry fee can submit high bids and win the auction causing the player who should have won to lose and pay the entry fee, whereas an irrational player in a reserve price auction cannot victimize others.

Ordinality and finiteness

First, we stress that, although our definition of a victim relied on the finiteness of the set of outcomes, it extends naturally to infinite sets of outcomes: we can simply consider the K^{th}

[14]To see this, consider two profiles R and R' with $a \in F(R)$. Suppose the least skilled candidate at R is b. For each member, the set of outcomes weakly less preferred to a at R that victimize at most $\omega = 1$ of the other members is simply $A \setminus \{b\}$. By assumption, b is not the highest skilled at R', and if a weakly rises in the ranking of each player among the set $A \setminus \{b\}$ going from R to R', then a remains the highest skilled at R' and $a \in F(R')$.

quantile instead of the least K preferred outcomes when defining a victim. Doing so would extend our results naturally to problems with an infinite set of outcomes.

Our definition of a victim uses ordinal preference information, making no use of cardinal information about the preferences of the players. This assumption may be violated in some settings of interest in mechanism design. However, many important subfields of mechanism design, such as voting and matching, rely on ordinal preferences (even though cardinal preferences may exist). To address safety in such settings, it is necessary to adopt a notion of safety that naturally accommodates ordinal preferences on finite sets of outcomes.

Our notion of safety can also be used as a basis for safety criteria in more general settings. For example, if utility functions are introduced, definition 2 can remain unchanged with the main change being in definition 3 of a victim; a victim can be defined to be a player who obtains an outcome leading to a utility level below a certain threshold. The utility threshold will, in general, be different for every player and be dependent on the precise utility function.

Finally, we note that it is common for policy makers to assess a mechanism's success (or failure) using ordinal welfare criteria very similar to the one used in this paper. One common criterion is the rank distribution, or how many players get their top choice, how many get their second choice, …, and how many get their last choice. Some examples are the San Francisco Unified School District, Teach for America, and Harvard Business School (Featherstone (2011)). Our safety criterion, along with our definition of a victim, is equivalent to using the rank distribution as a welfare measure–precisely what often occurs in many practical problems.

Thus, our safety criterion and notion of victimization address a large and important body of problems that are most naturally studied using ordinal welfare criteria. Our safety notion also opens new avenues for further research by providing a basis for developments of safety criteria in other settings.

Relative vs. Absolute Notions of Victimization

With our notion of safety, victims may be present at a "safe" strategy profile. An alternate definition, with serious drawbacks, can specify ω to be the maximum number of victims from a deviation *relative* to the number of victims in equilibrium; it is worth noting that such a definition can easily be accommodated by simply changing the notion of victimization. We do not use the "relative" definition precisely because we do not want many victims to be present at a "safe" strategy profile. To illustrate the drawbacks of a relative notion of victimization, consider two strategy profiles: m has no victims, and a deviation from m can lead to at most two victims, but m' has numerous victims, and a deviation from m' can lead to at most one more victim; a relative notion of safety may well consider m' to be "safer" than m, but our notion of safety would classify m' to be no safer than m. Also, a relative notion of safety

can allow a designer to make mechanisms seem more "safe" by simply letting the number of victims be high at all profiles, thus guaranteeing that profiles are not too bad, relatively.

Another possibility is to define a victim to be a player who does worse off (more generally: by a certain amount) from the equilibrium payoff, as in the t-immunity solution concept (Halpren (2008)). Again, a change in our notion of victimization can accommodate this definition, but this definition has some drawbacks. It does not allow for a clear statement of safety in terms of payoffs unless we know the payoffs of players at equilibrium, and we know which equilibria will be played (or we know that there is a unique equilibrium or a unique payoff across equilibria). Mechanisms generally have multiple equilibria with different payoffs. This definition may also artificially lead low-payoff equilibria to seem more safe. To illustrate some of the difficulties with this definition, consider a matching mechanism with two equilibria: one where all players get their top preferred outcome, but where a deviation can cause at most one non-deviator to get their second highest preferred outcome, and another equilibrium where all players get their least preferred outcome. A notion of victimization that focuses on payoffs of players *relative* to their equilibrium payoffs may consider the first equilibrium unsafe, but the second safe. In contrast, our notion of victimization would classify the second equilibrium to be no safer than the first. Additionally, using our notion of safety, if we, say, set $\alpha = \omega = K = 1$ and achieve implementation where all equilibria are (α,ω)-safe, then we can clearly say that no equilibrium will have more than one player obtaining their least preferred outcome, even if any other player deviates arbitrarily. A notion of safety relative to equilibrium payoffs cannot, in general, make such statement, because payoffs at equilibrium are generally unknown a-priori.

8.2 Safe Implementation Discussion

Is safe implementation easier than Nash implementation?

There is a common notion in the literature (e.g. Jackson (2001)) that a stronger (i.e. more refined) solution concept allows more social choice rules to be implemented, and thus makes the implementation problem "easier". If this were true then it would be easier to safely implement social choice rules than to implement them without regard for safety. This notion may be true in some cases, but not in general. Particularly, it is not true in the case of implementation in (α,ω)-safe equilibrium, a more refined solution concept than Nash equilibrium.

This can easily be seen using the simple social choice function in Example 2 in section 4. As the example shows, the social choice function satisfies Maskin monotonicity. It also satisfies no veto power and $n \geq 3$, so it is Nash implementable by Theorem 3 in Maskin (1999). However, the example shows that it does not satisfy (α,ω)-safe monotonicity, a necessary

condition for (α,ω)-safe implementation. Thus, a solution concept that is a Nash refinement does not necessarily allow more social choice rules to be implemented.

Existence of (α,ω)-Safe Implementable and Double Implementable Social Choice Rules

Theorem 8 is reminiscent of the Gibbard-Satterthwaite Theorem (Gibbard (1973) and Satterthwaite (1975)), which says that, under some conditions, a social choice function is strategy-proof if and only if it is dictatorial. Dictatorial social choice functions cannot be safe with a rich preference space (one player can always get any outcome they prefer), and strategy-proofness implies players have no incentive to deviate from reporting their true preferences. Theorem 8 says something similar: if a social choice function *is* safe (not dictatorial), then players will, in some preference profiles, have an incentive to deviate at any safe strategy profile of any implementing mechanism.

Though seemingly related, the Gibbard-Satterthwaite Theorem differs from Theorem 8 in important ways; the most important of these is perhaps the space of mechanisms each of the theorems operates on.[15] Theorem 8 says that no mechanism can implement a social choice function with safety guarantees in the space of *all* mechanisms (a very large space!). On the other hand, the Gibbard-Satterthwaite Theorem is only concerned with one particular mechanism: one where players report their preferences and the outcome is the social choice function at the reported preferences.

If dictatorial social choice functions are not safe with rich preference spaces, should we worry that the Gibbard-Satterthwaite Theorem implies non-existence of (α,ω)-safe implementable social choice rules? No; again, the Gibbard-Satterthwaite Theorem considers implementation using only one mechanism (truth-telling) and there may be other mechanisms that make implementation possible.

The necessary and sufficient conditions in Theorem 4 are not too restrictive. Simple social choice rules exist that are (α,ω)-safe implementable, such as the one in example 3: the example shows that (α,ω)-safe monotonicity holds trivially; the ω-safe outcome property holds (a victimizes no players at R and c victimizes no players at R'); and (α,ω)-exposure is satisfied (c victimizes all members in any group of two players at R and b victimizes all members in any group of two players at R'). Because $n = 3$ and $\alpha < \frac{n}{2}$, Theorem 4 implies that the social choice function in this example is (α,ω)-safe implementable, showing the existence of simple (α,ω)-safe implementable social choice rules. Also, the SCF in the hiring example of section 2.1 may or may not be dictatorial depending on the bonus premiums members assign candidates in their fields, but section 7 shows that this SCF is (α,ω)-safe implementable regardless of the bonus premiums of the members. For double implementation in safe and

[15]The results also concern different solution concepts: Gibbard-Satterthwaite Theorem is based on dominant strategies, but the solution concept in Theorem 8 is fundamentally based on Nash equilibrium.

Nash equilibria, section 7 shows that the SCF in our hiring example is double implementable with the appropriate preference restrictions, showing that natural social choice rules exist that are double implementable in safe and Nash equilibria.

9 Further Research and Conclusions

9.1 Further Research

Our victimization concept (definition 3) allowed us to apply our safety notion to the standard implementation problem, but, as mentioned before, our safety concept (definition 2) is much more general and can apply to virtually any mechanism design problem. A systematic study of safety (using definition 2) in various applications of mechanism design, perhaps with flexible context-dependent definitions of victimization, would be very useful in understanding safety more broadly. It would be interesting to examine and potentially improve the safety properties of common mechanisms in various applications, such as auctions, matching, voting, and market design.

In reality, society is often concerned with multiple safety objectives simultaneously, such as protection from irrational and dangerous individuals and larger, more organized groups deciding to harm others (e.g. with discriminatory policies). Our notion of safety imposes just one restriction on the victims function: $V_m(\alpha) \leq \omega$, where m is the equilibrium profile.[16] This can be extended in various ways: more restrictions at multiple points of the victims function can reflect more complicated safety requirements; the victims function at non-equilibrium profiles may also be restricted to improve robustness; and finally, multiple definitions of victimization (e.g. multiple K's or fundamentally different definitions of victimizations) can be used, leading to multiple victims functions, each with its own set of α's and ω's.

Furthermore, in reality we may not always be able to eliminate disasters, but we can potentially make them unlikely. Adopting a probabilistic approach to safety would be a fruitful area for future research. Our victims function is concerned only with the worst case scenario at a given α, but a probabilistic approach can examine the whole distribution of the number of victims.

In this paper we studied safety in conjunction with the Nash equilibrium as a solution concept. An interesting extension would be to add safety requirements to an arbitrary solution concept. In particular, it would be interesting to impose safety restrictions in settings where players learn to play.

[16] If $\omega < n - \alpha$ then $V_m(\alpha) \leq \omega$ also implies that $V(\beta) \leq \omega$ for any $\beta < \alpha$.

9.2 Conclusions

For system designers and policy makers safety is of utmost importance. Despite this fact, many mechanisms in the implementation theory and mechanism design literature have *no* safety properties: for most mechanisms in the literature, an unexpected deviation by just one player deviation can cause widespread disasters. Introducing safety requirements can change the set of implementable social choice rules, so it is important to address safety rigorously and examine the constraints implied by safety.

This paper bridges the gap between safety and the mechanism design and implementation theory literature. We introduced safety as robustness of the welfare levels of the players to actions of others, and we characterized social choice rules that can be implemented in a safe way. We introduced three notions of safe implementation: (1) strong (α,ω)-safe implementation is Nash implementation with mechanisms where all Nash equilibria are (α,ω)-safe; (2) (α,ω)-safe implementation, which is a weaker notion of implementation that allows for some undesirable Nash equilibria; and (3) Double implementation in (α,ω)-safe and Nash equilibria, which maintains safe implementation and ensures that all Nash equilibria lead to desirable outcomes. We provided necessary conditions for (α,ω)-safe implementation, and we showed that they are sufficient under some conditions on preferences. We also provided conditions that are both necessary and sufficient for double implementation when transfers are allowed. Double implementation leads to strong safe implementation with mild behavioral assumptions unrelated to safety. We proved a "no guarantees" result showing the impossibility of implementation in $(n-1,0)$-safe equilibria in a wide range of environments. Finally, we discussed the connection between safety, freedom, and efficiency showing that, in general, safety restricts players' positive freedom and reduces efficiency.

APPENDIX

Proof of Theorem 1

Let (M,g) be a mechanism that strongly (α,ω)-safe implements F. To prove this theorem, we will show that if condition (1) does not hold, then condition (2) holds. Suppose $a \in F(R)$ and condition (1) doesn't hold so that $\exists R^{'}$ with $a \notin F(R^{'})$. Because F is strongly (α,ω)-safe implementable, there is a Nash equilibrium at R, say m, such that $g(m) = a$. Because $a \notin F(R^{'})$, m cannot be a NE at $R^{'}$. Hence, at least one player, say player i, has a profitable deviation from m to obtain some other outcome, say b, at $R^{'}$ and thus, $bP_i^{'}a$. At R, player i must weakly prefer a to b, otherwise obtaining b would be profitable for player i at R and m would not be a NE at R. Hence, aR_ib. Now consider another profile, $R^{''}$, where a rises in all players' ranks relative to R. m must then be a NE at $R^{''}$. Because every NE is (α,ω)-safe and

because player i can obtain b by deviating from m, b cannot victimize more than ω players in $I \backslash \{i\}$ at R''. Hence, condition (2) holds.

Proof of Corollary 1

To show a contradiction, suppose (M, g) is a mechanism that strongly (α, ω)-safe implements F on $\mathscr{R}$ and $\mathscr{R}$ contains all possible strict preferences. Let $a \in F(R)$ for some $R \in \mathscr{R}$. We will first show that condition (1) of Theorem 1 holds by showing that condition (2) of that theorem does not hold. Suppose $a \notin F(R')$ for some $R' \in \mathscr{R}$. Because the set of admissible preference profiles contain all possible preference profiles, for every pair of player and outcome $(i, b) \in I \times A$, there is a preference profile where a weakly rises in every player's preferences relative to R and b is the least preferred by all players, violating condition (2). Hence, condition (1) holds and, because a was arbitrary, F is a constant social choice rule. Consider a profile, $R^* \in \mathscr{R}$, where a is strictly the least preferred outcome by all players. By condition (1) of Theorem 1, $a \in F(R^*)$. By strong (α, ω)-safe implementability, there is a NE at R^*, say m^*, such that $g(m^*) = a$. However, a victimizes all players so m^* cannot be safe, leading to a contradiction.

Proof of Theorem 2

The necessity of the ω-safe outcome property is trivial as it follows directly from the (α, ω)-SE definition. To show the necessity of (α, ω)-safe monotonicity, suppose that F is implementable in (α, ω)-safe equilibrium by mechanism (M, g) and suppose that R and R' are two preference profiles such that $a \in F(R)$ and conditions (1) and (2) of the (α, ω)-safe monotonicity property hold. By implementability, there must be a (α, ω)-safe equilibrium $m \in M$ at R such that $g(m) = a$. For any $i \in I$ let $O_i(m) \subseteq A$ be the set of outcomes that player i can obtain by a (potential) deviation from m. Because m is a (α, ω)-safe equilibrium it must be that for any $y \in O_i(m)$, $aR_i y$ and y victimizes at most ω of the players in set $I \backslash \{i\}$ under R. By condition (1) of (α, ω)-safe monotonicity, this must also hold under R'. Additionally, (in the case of $\alpha \geq 2$), for any subset of players $B \subseteq I$ of size α, let $O_B(m)$ be the attainable set by players in B from m. Because m is a (α, ω)-safe equilibrium, it must be that every element in $O_B(m)$ victimizes at most ω of the players in $I \backslash B$ under R. Condition (2) of (α, ω)-safe monotonicity implies that the same holds for R'. Hence, we have shown that, under R', the profile m is such that no player can gain by deviating and no group of $\alpha \geq 1$ players can victimize more than ω of the others by deviating. Therefore, m is a (α, ω)-safe equilibrium at R' and $a \in F(R')$.

Proof of Theorem 3

Suppose that F is implementable in (α,ω)-safe equilibrium by mechanism $\Gamma=(M,g)$ and suppose that $\alpha \geq \frac{n}{2}$. Let $\mathscr{R}$ be the set of preference profiles and for each $R^j \in \mathscr{R}$ and $g^j \subseteq I$ let $H(g^j,R^j) \subseteq A$ be the set of outcomes that cause ω or less victims in g^j under profile R^j. Suppose, by way of contradiction, that the set of preference profiles, $\mathscr{R}$, violates the (α,ω,n)-similarity property. Then, there exists a set $\{(g^1,R^1),\ldots,(g^q,R^q)\}$ with $q=\left\lfloor \frac{n}{n-\alpha} \right\rfloor$, $|g^i| \geq n-\alpha$, and $g^i \cap g^k = \varnothing$, $i,k \in \{1,\ldots,q\}$, and where each outcome in A causes more than ω victims in g^j at R^j for some $j \in \{1,\ldots,q\}$; that is to say, $\bigcap_{j\in\{1,2,\ldots,q\}} H(g^j,R^j)$ is empty. For each $j \in \{1,\ldots,q\}$ let $SE^{\Gamma}(R^j)$ be the set of (α,ω)-SE of Γ under R^j. Suppose m is any profile where $n-\alpha$ or more players g^j follow the strategies in some $m^{'} \in SE^{\Gamma}(R^j)$. Then players in $I \backslash g^j$, a set of α or less players, cannot victimize more than ω players in g^j. Thus, $g(m) \in H(g^j,R^j)$ by the fact that $m^{'}$ is a (α,ω)-SE at R^j. By the definition of q and because $\alpha \geq \frac{n}{2}$, there exists a profile, m^*, and q different groups of players of size at least $n-\alpha$, call them $\{g^1,\ldots,g^q\}$, such that for each $j \in \{1,\ldots,q\}$, $m^*_{g^j} = m^{'}_{g^j}$ for some $m^{'} \in SE^{\Gamma}(R^j)$. Hence, it must be that $g(m^*) \in \bigcap_{j\in\{1,\ldots,q\}} H(g^j,R^j)$. However, $\bigcap_{j\in\{1,2,\ldots,q\}} H(g^j,R^j)$ is empty by assumption, leading to a contradiction.

Proof of Theorem 4

Suppose first that $\alpha < \frac{n}{2}$ and the conditions of the Theorem are satisfied. Let $\Gamma=(\times_{i\in I} M_i, g)$ be the implementing mechanism, where the message space for each player $i \in I$ is $M_i = A \times \mathscr{R}^* \times \{1,\ldots,n\}$ and $\mathscr{R}^*$ is the set of preference profiles. A typical message is $m_i=(a^i,R^i,z^i)$. The outcome function, g, is determined as follows:

Rule 1: If $(a^i,R^i,z^i)=(a,R,z)$ for all $i \in I$ and $a \in F(R)$ then $g(m)=a$.

Rule 2: Suppose there exists i such that $(a^j,R^j,z^j)=(a,R,z)$ for all $j \neq i$, $a \in F(R)$, and $(a^i,R^i,z^i) \neq (a,R,z)$. Then $g(m)=a^i$ if aR_ia^i and if a^i does not victimize more than ω players in the set $I\backslash\{i\}$ under R; otherwise, $g(m)=a$.

Rule 3: Suppose $\alpha \geq 2$ and there exists a set of players $B \subseteq I$ with $1<|B| \leq \alpha$ such that $(a^j,R^j,z^j)=(a,R,z)$ for all $j \notin B$, $a \in F(R)$, and $(a^i,R^i,z^i) \neq (a,R,z)$ for all $i \in B$. Let $B^{'}$ be the set of players in B who report an outcome that does not victimize more than ω players in $I\backslash B$ under R. If $B^{'}$ is not empty, then sort the players in $B^{'}$ from lowest to highest according their index, rename player $i \in B^{'}$ by the new sorted order $q_i \in \{1,\ldots,|B^{'}|\}$, and let $g(m)=a^{q_i^*}$, where $q_i^* = 1+((\sum_{i\in B^{'}} z^i) \mod |B^{'}|)$ is the q_i^*'th player in $B^{'}$; otherwise $g(m)=a$.

Rule 4: In all other cases, $g(m)=a^j$ for $j=1+((\sum_{i\in I} z^i) \mod n)$.

Let R be the true preference profile and suppose $a \in F(R)$. We first show that there exists a safe equilibrium m of the mechanism Γ such that $g(m)=a$. Let m be given by $m_i=(a,R,1)$.

By rule 2, any unilateral deviation from m cannot result in a more preferred outcome under R for the deviator, so m is a Nash equilibrium. Also under rule 2, a unilateral deviation from m by an arbitrary player i cannot cause more than ω players in $I\backslash\{i\}$ to be victimized under R. A multilateral deviation from m by an arbitrary set B of α or less players falls under rule 3, which guarantees that the outcome implemented will not victimize more than ω players in $I\backslash B$ under R. Finally, a does not victimize more than ω players under R because F satisfies the ω-safe outcome property. Hence, m is a (α,ω)-safe equilibrium.

Next, I show that if m is a (α,ω)-SE of Γ then $g(m)\in F(R)$. Because of the assumption that R satisfies (α,ω)-exposure, no (α,ω)-SE of Γ can fall under rules 2-4. If there were a (α,ω)-SE under rules 2-4 it would be possible for a group, B, of α players to attain any outcome in A via rule 4.[17] By assumption, R satisfies (α,ω)-exposure so the set of outcomes that cause more than ω players in $I\backslash B$ to be victims is nonempty. This implies that a group of α players can victimize more than ω of the rest, violating the definition of the (α,ω)-SE. Hence, the only (α,ω)-safe equilibria of Γ fall under rule 1.

Suppose m is a (α,ω)-SE under rule 1 such that $m_i=(a^{'},R^{'},z^{'})$ for all $i\in I$. For an arbitrary player i, let $O_i(m)\subseteq A$ be the set composed of $a^{'}$ and all the outcomes that player i can obtain by deviating from m. By rule 2 and the ω-safe outcome property, if $y\in O_i(m)$ then $a^{'}R_i^{'}y$ and y must not victimize more than ω players in $I\backslash\{i\}$ under $R^{'}$. However, because m is a (α,ω)-SE under the true preference profile, R, it must be that $a^{'}R_iy$ and y cannot victimize more than ω players in $I\backslash\{i\}$ under R. Additionally, rule 2 implies that $O_i(m)$ contains *all* the outcomes that are both weakly less preferred to $a^{'}$ under $R^{'}$ by player i and victimize at most ω of the players in $I\backslash\{i\}$ under $R^{'}$. Hence, the first condition of (α,ω)-safe monotonicity is satisfied.

Suppose $\alpha\geq 2$. Then for an arbitrary set of players, B, of size α, let $O_B(m)$ be set composed of $a^{'}$ and all the outcomes that players in B can obtain by a (potentially multilateral) deviation from m. By rules 2 and 3 and the ω-safe outcome property, if $y\in O_B(m)$ then y victimizes at most ω players in $I\backslash B$ under $R^{'}$. However, m is a (α,ω)-SE under R, so it must be that y victimizes at most ω players in $I\backslash B$ under R. Because $O_B(m)$ contains, via rules 2 and 3, *all* elements that victimize at most ω players in $I\backslash B$ under $R^{'}$, the second condition of (α,ω)-safe monotonicity is satisfied.

By (α,ω)-safe monotonicity, $a^{'}\in F(R)$, concluding the proof for the case of $\alpha<\frac{n}{2}$.

If $\frac{n}{2}\leq\alpha<n-1$ the proof remains unchanged except that the set of possible profiles, $\mathscr{R}^*$, is assumed to be (α,ω,n)-similar and rule 3 of Γ is modified to be:

> **Rule 3:** Suppose rule 2 does not apply and suppose there are p *consensus groups*, $B_1,\ldots,B_p\subseteq I$, each of size $n-\alpha$ or more players with everyone in B_1 reporting

[17] Rule 4 is reached when there are more than α *different* strategies reported by the players. Note that it is enough for a strategy to differ only on the R dimension in order to be counted "different".

$(\bar{a}^1,\bar{R}^1,\bar{z}^1)$, everyone in B_2 reporting $(\bar{a}^2,\bar{R}^2,\bar{z}^2)$, ..., and everyone in B_p reporting $(\bar{a}^p,\bar{R}^p,\bar{z}^p)$. Suppose also that $(\bar{a}^1,\bar{R}^1,\bar{z}^1) \neq (\bar{a}^2,\bar{R}^2,\bar{z}^2) \neq \ldots \neq (\bar{a}^p,\bar{R}^p,\bar{z}^p)$, and if i is a player not in any of the consensus groups then $(a^i,R^i,z^i) \neq (\bar{a}^h,\bar{R}^h,\bar{z}^h)$ for $h \in \{1,\ldots,p\}$.

- Case of $p = 1$. Set $g(m) = a^q$ for $q = \max\{j \in I \backslash B_1\}$ if no element in the set $\{a^j : j \in I \backslash B_1\}$ victimizes more than ω players in set B_1 under profile $\bar{R}^1$; otherwise $g(m) = \bar{a}^1$.
- Case of $p > 1$. Set $g(m) = b$ for some b that does not victimize more than ω players in B_j under $\bar{R}^j$ for all $j \in \{1,\ldots,p\}$. Such a b exists because of the assumption that $\mathscr{R}^*$ is (α,ω,n)-similar.

Note that because $\frac{n}{2} \leq \alpha < n-1$, no (α,ω)-SE can fall under the modified rule 3. It is always true that α players can deviate from the modified rule 3 and reach rule 4. Also note that, through case $p = 1$ of the modified rule 3, a set of deviators from rule 1 can attain all outcomes that do not victimize more than ω players of the rest. This is needed to satisfy (α,ω)-safe monotonicity.

Proof of Theorem 5

Suppose F is implementable in both (α,ω)-safe and Nash equilibria. If $a \in F(R)$ and safe implementation holds, then there is a safe equilibrium, call it m, at R with $g(m) = a$. If player i can deviate from m and obtain outcome b then aR_ib and b victimizes at most ω players in $I\backslash\{i\}$, because m is a safe equilibrium. If all such outcomes are weakly less preferred than a for player i under preference profile R', and if this holds for all players (as ω-safe-Nash monotonicity states), then m is a Nash equilibrium at R' and $a \in F(R')$ by Nash implementation.

Proof of Theorem 6

Consider the mechanism used in the proof of Theorem 4 with the following transfer scheme:

Transfer Scheme 1

Rule 1: All transfers are zero under rule 1.

Rule 2: Suppose all players agree on (a,R,z) with $a \in F(R)$, except for one player i who reports $(a',R',z') \neq (a,R,z)$. If $R' \neq R$ and $z' \neq z$ then all transfers are zero; otherwise $t_i = -L$ and $t_j = 0$ for $j \neq i$.

Rule 3: Under this rule $\alpha \geq 2$ and all players agree on (a,R,z), except a set of players B who submit different reports and $1 < |B| \leq \alpha$. Set the transfer of all players in B to $-L$ and all others to zero unless the following is true: $|B| = 2$, call them players i and

j, with player i reporting (a',R',z') and player j reporting (a'',R'',z''), with $R' \neq R$, $z' \neq z$, $R'' = R$, and $z'' = z'$, then set $t_i = -L$, $t_j = L$, and all other transfers to zero.

Rule 4: Set $t_i = -L\sum_{j\neq i}\mathbf{1}\{z^j = 1 + (z^i \mod n)\} + L\sum_{j\neq i}\mathbf{1}\{z^i = 1 + (z^j \mod n)\}$.

We will prove that the mechanism used in Theorem 4, along with the transfers in Transfer Scheme 1, double implement F. Let R be the true preference profile and suppose $a \in F(R)$.

First note that, following the same argument in the proof of Theorem 4, the strategy profile where all players report $(a,R,1)$ is a (α,ω)-safe equilibrium with outcome a. Thus, all outcomes in F can be obtained via a (α,ω)-safe equilibrium (i.e. a Nash equilibrium) of the mechanism with transfers.

Second, we show that if m is a Nash equilibrium of the mechanism with transfers then $g(m) \in F(R)$ (this also implies that all (α,ω)-safe equilibria lead to outcomes in $F(R)$).

Suppose m is a Nash equilibrium under rule 1 with every player reporting (a',R',z'). For an arbitrary player i, let $O_i(m) \subseteq A$ be the set composed of a' and all the outcomes that player i can obtain by deviating from m. By rule 2 and the ω-safe outcome property, if $b \in O_i(m)$ then $a'R'_i b$ and b does not victimize more than ω players in $I\backslash\{i\}$ under R'. However, because m is a Nash equilibrium under the true preference profile R, and because player i can obtain any outcome in $O_i(m)$ without getting fined (by reporting a different preference profile and integer than other players), it must be that $a'R_i b$. Additionally, rule 2 implies that $O_i(m)$ contains all the outcomes that are both weakly less preferred to a' under R' by player i and victimize at most ω of the players in $I\backslash\{i\}$ under R'. Hence, $a' \in F(R)$ by (α,ω)-safe-Nash monotonicity.

Next, we show that no Nash equilibria exist under rules 2,3, and 4. Suppose m is a Nash equilibrium under rule 2 with player i reporting (a',R',z') and all others reporting $(\overline{a},\overline{R},\overline{z})$. Player i cannot be fined at a Nash equilibrium under rule 2, because he can report $R' \neq \overline{R}$ and $z' \neq \overline{z}$ and avoid the fine without affecting the outcome chosen. Thus, it must be true that $R' \neq \overline{R}$ and $z' \neq \overline{z}$. However, any other player can deviate and report $(\overline{a},\overline{R},z')$ and obtain a transfer L from player i, via rule 3, and this deviation will be profitable, contradicting m being a Nash equilibrium.

Suppose m is a Nash equilibrium under rule 3 with all players agreeing on $(\overline{a},\overline{R},\overline{z})$, except a set of players B with different reports. At least one player in B has transfer $-L$, and any such player has an incentive to deviate to $(\overline{a},\overline{R},\overline{z})$ and avoid the fine, contradicting m being a Nash equilibrium.

Finally, suppose m is a Nash equilibrium under rule 4. For the rest of the proof we assume modular n arithmetic, so that if $z = n$ then $z+1 = 1$, $z+2 = 2$, etc... There must be an integer that no player is reporting, otherwise each integer in $\{1,\ldots,n\}$ is being reported by exactly one player and all players would have an incentive to deviate (e.g. the player reporting integer n can deviate to reporting $n-1$ and have transfer L instead of 0). Let z^* be the

smallest element in $\{1,\ldots,n\}$ such that no player reports z^* and at least one player reports z^*-1. Because no player reports z^*, it must be that no player has a negative transfer, otherwise a player with negative transfer can deviate to reporting z^*-1 and pay no fine. Because the transfers among the agents are zero-sum, each agent's transfer is zero. At least 2 players must be reporting z^*+1, otherwise at least one player (possibly one reporting z^*+1) would have an incentive to deviate and report z^* to obtain a positive transfer. Because players reporting z^*+1 have zero net transfers, and because no players report z^*, it must be that no players report z^*+2. Following the same argument, at least two players report z^*+3 and no players report z^*+4, at least two players report z^*+5 and no players report z^*+6, etc... Such reporting behavior is only possible if exactly two players are reporting each of z^*+1, z^*+3, z^*+4, etc... But then a player reporting z^*+3 has an incentive to deviate and report z^*+2, obtaining a transfer of L. This leads to a contradiction and m cannot be a Nash equilibrium.

Proof of Corollary 2

Suppose the preference profile is R. In the proof of Theorem 6, we showed the following: for any outcome $a \in F(R)$, there exists a safe equilibrium, m, with $g(m)=a$; for any Nash equilibrium under rule 1 where all players report (a',R',z'), $a' \in F(R)$; and that there are no Nash equilibria under rules 2, 3, or 4. Thus, any unsafe Nash equilibrium must fall under rule 1, where all players agree on a report. If the preference profile being reported by all players is R under rule 1, then the message profile is safe by the design of the mechanism (in the same way that we showed the profile where each player reports $(a,R,1)$ is safe in the proof of Theorem 4). Hence, it must be that all players report a false preference profile at any unsafe Nash equilibrium.

Suppose m is an unsafe Nash equilibrium. Double implementation implies that $g(m) \in F(R)$. Thus, the (truthful) profile where each player reports $(g(m),R,1)$ falls under rule 1, is safe by the design of the mechanism, and yields the same outcome as m.

Proof of Theorem 7

Suppose F is double implementable in (α,ω)-safe and Nash equilibria, but ω-safe-Nash monotonicity fails. Thus, there are two preference profiles R and R' with: $a \in F(R)$, and, for all $i \in I$, aR_ib and b victimizes at most ω players in $I\backslash\{i\}$ implies aR'_ib, but $a \notin F(R')$. There must be a safe equilibrium at R with outcome a, call it m. $a \notin F(R')$ by assumption, so m cannot be a Nash equilibrium at R', otherwise $g(m)=a$ must be in $F(R')$ by Nash implementation. Thus, some player, call it i, has a profitable deviation from m to another profile, call it m', at R'. The outcome at m' must be different than a, otherwise the transfer

required to induce a deviation from m to m' at R' would also induce the same deviation at R and m would not be a Nash equilibrium at R; let the outcome at m' be b. Because m is a safe equilibrium at R, aR_ib and b victimizes at most ω players in $i\backslash\{i\}$ at R, and – by our assumption about the failure of ω-safe-Nash monotonicity between R and R' – aR'_ib; thus, $(a,0)R_i(b,0)$ and $(a,0)R'_i(b,0)$. However, player i must have an incentive to deviate to m' at R', and this is only possible if the designer can set transfers $t(m),t(m') \in \mathbb{R}$ such that: (1) $(a,t(m))R_i(b,t(m'))$, because m is a Nash equilibrium at R; and (2) $(b,t(m'))P_i(a,t(m))$, because player i has an incentive to deviate from m to m' at R'. To set such transfers, or determine that none exist, is only possible if the designer has cardinal information about preferences for player i and outcomes $\{a,b\}$ at preference profiles $\{R,R'\}$ as defined in definition 12.

Proof of Lemma 1

For the purpose of deriving a contradiction assume that there exists a mechanism, $(\times_{i\in I}M_i,g)$, that implements a social choice function f in a $(n-1,0)$-SE and that for some i, $\exists D \subseteq A$ such that $|D| = K$ and $\forall m_i \in M_i$ $\exists m_{-i} \in M_{-i}$ with $g(m_i,m_{-i}) \in D$. Since implementation must hold for every possible preference profile, consider the preference profile where D is player i's least favorite K outcomes. This profile is one of the possible ones because the set of preference profiles contains all strict preferences. Let m be a $(n-1,0)$-SE. By assumption, $\exists m'_{-i} \in M_{-i}$ such that $g(m_i,m'_{-i}) \in D$ violating the definition of a $(n-1,0)$-SE.

Proof of Theorem 8

Suppose $x \leq nK + K\left\lfloor\frac{n}{2}\right\rfloor$ and let f be any SCF.

Case 1: $x \leq nK$. Let $R \in \mathscr{R}$ be the preference profile where the least preferred K elements for player 1 are elements $a_1,...,a_K$, the least preferred for player 2 are $a_{K+1},...,a_{2K}$, etc... Since $x \leq nK$ then at some point each element in A will be among the least preferred K ones for some $i \in I$. Hence, any element picked by f in this case will result in at least 1 victim and f is not implementable in a $(n-1,0)$-SE.

Case 2:[18] $nK < x \leq nK + K\left\lfloor\frac{n}{2}\right\rfloor$. By Lemma 1, any mechanism, $(\times_{i\in I}M_i,g)$, that implements a social choice function in this case must allow every player to rule out their least favorite K outcomes. Furthermore, in a $(n-1,0)$-SE every player must be following a strategy that rules out their least favorite K outcomes because if not then $\exists m_{-i} \in M_{-i}$ so that the outcome is among the K least favorite for player i violating the definition of a $(n-1,0)$-SE. Let $R \in \mathscr{R}$ be any preference profile with the following properties:

- The second least favorite K outcomes for player i are $\{a_{(i-1)K+1},\ldots,a_{iK}\}$ $\forall i \in I$.

[18] See example 4

- The least favorite K outcomes for players 1 and 2 are $\{a_{nK+1},\ldots,a_{nK+K}\}$, for players 3 and 4 are $\{a_{nK+K+1},\ldots,a_{nK+2K}\}$, etc..., until a_x is among the least favorite K outcomes for some player, call it player j, then sequentially use $\{a_1,\ldots,a_{K-1}\}$ to fill the rest of the K least favorite outcomes for player j. The least favorite K outcomes for players $j+1,\ldots,n$ are the same as player j's.

Hence, R, is as follows:

R_1	R_2	R_3	R_4	$\ldots$	R_j	R_{j+1}	$\ldots$	R_n
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_1	a_{K+1}	a_{2K+1}	a_{3K+1}	$\ldots$	$a_{(j-1)K+1}$	a_{jK+1}	$\ldots$	$a_{(n-1)K+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_K	a_{2K}	a_{3K}	a_{4K}	$\ldots$	a_{jK}	$a_{(j+1)K}$	$\ldots$	a_{nK}
a_{nK+1}	a_{nK+1}	a_{nK+K+1}	a_{nK+K+1}	$\ldots$	$a_{nK+cK+1}$	$a_{nK+cK+1}$	$\ldots$	$a_{nK+cK+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	a_x	a_x	$\ldots$	a_x
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	a_1	a_1	$\vdots$	a_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_{nK+K}	a_{nK+K}	a_{nK+2K}	a_{nK+2K}	$\ldots$	a_y	a_y	$\ldots$	a_y

where c is the integer quotient from dividing $j-1$ by 2 (and ignoring the remainder), and $y \in \{1,\ldots,K-1\}$. The set of outcomes below the dashed line are the least favorite K outcomes for each player.

Since we know that every player must pick a strategy that rules out their least favorite K outcomes, then any outcome picked by the social choice function will be in $\{a_1,\ldots,a_{nK}\}$. Hence, at any equilibrium $(m_1,...,m_n)$ of the game induced by the mechanism under R, the outcome is the among $\{a_{(i-1)K+1},\ldots,a_{iK}\}$ for some player i. However, since the least favorite K outcomes of player i are being ruled out by at least one other player ($x \leq nK + K\left\lfloor\frac{n}{2}\right\rfloor$), player i has an incentive to deviate from this equilibrium and rule out $\{a_{(i-1)K+1},\ldots,a_{iK}\}$ in order to get a choice that is more preferred for player i under R_i. By Lemma 1 we know that player i has a strategy that rules out $\{a_{(i-1)K+1},\ldots,a_{iK}\}$. Hence, $(m_1,...,m_n)$ is not a $(n-1,0)$-SE because it is not a Nash equilibrium.

Discussion of Section 6.2

Suppose first that $\alpha = 1$, the preference profile is R and consider a mechanism that (α,ω)-safe implements a SCR F. A strategy profile, m, of the mechanism is a (α,ω)-safe equilibrium if and only if $O_i(m) \subseteq L_i(a,R_i) \cap G_i(R)$ for all i, where a is the outcome at m. It follows that if at another preference profile, R', we have $L_i(a,R_i) \cap G_i(R) \subseteq L_i(a,R_i') \cap G_i(R')$ then it must

be true that $O_i(m) \subseteq L_i(a,R_i^{'}) \cap G_i(R^{'})$, m is a (α,ω)-SE at $R^{'}$, and $a \in F(R^{'})$. This is precisely the statement of (α,ω)-safe monotonicity when $\alpha = 1$. Figure 1 illustrates what must happen for each player between preference profiles R and $R^{'}$ for (α,ω)-safe monotonicity to impose the restriction that $a \in F(R^{'})$.

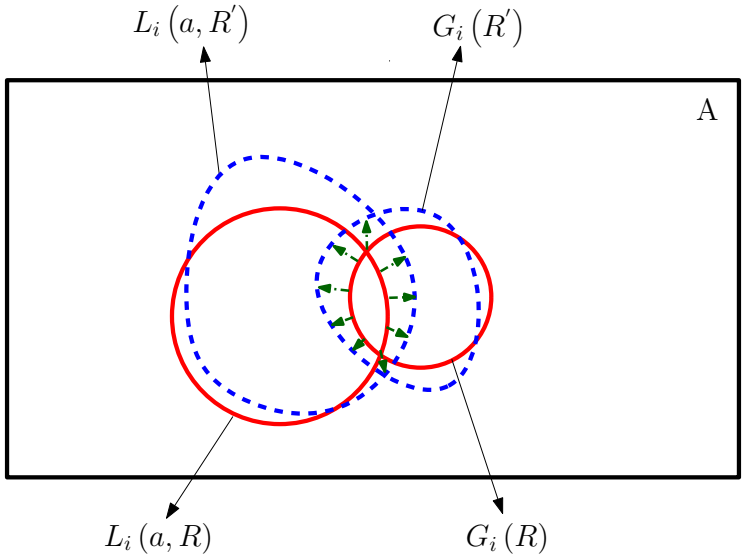

FIGURE 1. Condition (1) of (α,ω)-Safe Monotonicity

The figure shows that the intersections of the sets $L_i(a,\cdot)$ and $G_i(\cdot)$ must (weakly) increase moving from R to $R^{'}$. Note, as the figure illustrates, the lack of restrictions on the relationship between the lower contour sets $L_i(a,R)$ and $L_i(a,R^{'})$; this contrasts with Maskin monotonicity where the only restriction is on the inclusion of $L_i(a,R)$ in $L_i(a,R^{'})$

Figure 1 also illustrates why (α,ω)-safe monotonicity does not imply Maskin monotonicity and vise versa: $L_i(a,R) \cap G_i(R) \subseteq L_i(a,R^{'}) \cap G_i(R^{'})$ (condition (1) of (α,ω)-safe monotonicity) does not imply $L_i(a,R) \subseteq L_i(a,R^{'})$ (the Maskin monotonicity condition). Similarly, if $L_i(a,R) \subseteq L_i(a,R^{'})$ were true it would not imply $L_i(a,R) \cap G_i(R) \subseteq L_i(a,R^{'}) \cap G_i(R^{'})$. However, it is certainly true that both conditions may hold at the same time.

When $\alpha \geq 2$ then a second condition must be added to the definition of (α,ω)-safe monotonicity to take into account the possible presence of more than one deviator. However, the first condition, which is described above and illustrated in Figure 1, must hold for (α,ω)-safe monotonicity to impose a restriction on the SCR regardless of α.

REFERENCES

[1] Aumann, R. (1960), "Acceptable points in games of perfect information", *Pacific Journal of Mathematics*, **10**, 381-417

[2] Bergemann, D. and S. Morris. (2005), "Robust Mechanism Design", *Econometrica*, **73**, 1771-1813

[3] Bernheim, D., D. Peleg, and M. Whinston. (1987), "Coalition-Proof Nash Equilibria I. Concepts", *Journal of Economic Theory*, **42**, 1-12

[4] Cabrales, A., and R. Serrano. (2011), "Implementation in adaptive better-response dynamics: Towards a general theory of bounded rationality in mechanisms", *Games and Economic Behavior*, **73**, 360-374

[5] Eliaz, K. (2002), "Fault Tolerant Implementation", *Review of Economic Studies*, **69**, 589-610

[6] Featherstone, C. (2011), "Rank Efficiency: Investigating a Widespread Ordinal Welfare Criterion", Job Market Paper

[7] Gibbard, A. (1973). "Manipulation of voting schemes: A general result." *Econometrica*, **41**, 587-601

[8] Halpren, J. (2008). "Beyond Nash Equilibrium: Solution Concepts for the 21st Century", *Proceedings of the Twenty-Seventh Annual ACM Symposium on Principles of Distributed Computing*

[9] Harsanyi, J., and R. Selten. (1988), *A General Theory of Equilibrium Selection in Games*. MIT Press.

[10] Jackson, M. (2001), "A Crash Course in Implementation Theory", *Social Choice and Welfare*, **18**, 655-708

[11] Maskin, E. (1999), "Nash Equilibrium and Welfare Optimality", *Review of Economic Studies*, **66**, 23-38

[12] Muller, E. , and M. Satterthwaite (1977), "The Equivalence of Strong Positive Association and Strategy-proofness", *Journal of Economic Theory*, **14**, 412-418

[13] Satterthwaite, M. (1975). "Strategy proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare theorems." *Journal of Economic Theory*, **10**, 187-217

[14] Selten, R. (1975), "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games", *International Journal of Game Theory*, **4**, 25-55

[15] Wen-Tsun, W., and J. Jia-He. (1962), "Essential equilibrium points of n-person non-cooperative games", *Scientia Sinica*, **11**, 1307-1322