Safe Implementation

BSE Working Paper 1363

September 2022 (Revised: May 2023)

Malachy James Gavan, Antonio Penta

Safe Implementation[*]

Malachy James Gavan[†]
UPF and BSE

Antonio Penta[‡]
ICREA-UPF, BSE and TSE

May 9, 2023

Abstract

We introduce *Safe Implementation*, a new framework for implementation theory that adds to the standard requirements the restriction that deviations from the baseline solution concept induce outcomes that are *acceptable*. The primitives of Safe Implementation therefore include both a Social Choice Correspondence, as standard, and an Acceptability Correspondence, each mapping every state of the world to a subset of allocations. This framework generalizes standard notions of implementation, and can accommodate a variety of questions, including robustness concerns with respect to mistakes in play, model misspecification, behavioral considerations, state-dependent feasibility restrictions, limited commitment, etc.

We provide results both for general solution concepts and for the case in which agents' interaction is modelled by Nash Equilibrium. In the latter case, we identify necessary and sufficient conditions (namely, *Comonotonicity* and *safety-no veto*) that restrict the joint behavior of the Social Choice and Acceptability Correspondences. These conditions are more stringent than Maskin's (1977), but coincide with them when the safety requirements are vacuous. We also show that these conditions are quite permissive in important economic applications, such as environments with single-crossing preferences and in problems of efficient allocation of indivisible goods, but also that Safe Implementation can be very demanding in environments with 'rich' preferences, regardless of the underlying solution concept.

Keywords: Comonotonicity – mechanism design – implementation – robustness – resilience – safe implementation – safety no-veto

JEL Codes: C72; D82.

1 Introduction

Since Maskin (1977, 1999)'s seminal work, implementation theory has played a central role in developing our understanding of market mechanisms, institutions, and their foundations. The theory starts out by specifying a set of agents, a set of states – that pin down agents' preferences over the feasible allocations – and a Social Choice Correspondence (SCC) that specifies, for each state,

[*] We are grateful to Larbi Alaoui, Mehmet Barlo, Aygun Dalkiran, Pia Ennuschat, Alexander Frug, Bard Harstad and William Sandholm for their comments. The BSE acknowledges the financial support of the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S). Antonio Penta acknowledges the financial support of the European Research Council (ERC), ERC Starting Grant #759424.

[†] *email:* malachy.gavan@upf.edu

[‡] *email:* antonio.penta@upf.edu

the set of allocations that the designer wishes to induce. While commonly known by the agents, the state of nature is unknown to the designer, and hence in order to choose the allocation the designer must rely on agents' reports. The main objective of the theory is to study the conditions under which it is possible to specify a mechanism in which, at every state, the allocations selected by the SCC are sustained as the result of agents' strategic interaction. The latter is suitably modeled via game theoretic solution concepts, each giving rise to different notions of implementation.[1]

In its baseline form, the theory imposes no restriction on the mechanisms that may achieve implementation, nor on the outcomes that may arise from agents' deviations, beyond the fact that they provide the right incentives.[2] For instance, a standard argument in the literature is the idea that incentives may sometimes be easily provided by applying a "shoot the deviator" kind of logic. In practice, though, the designer does not always have this freedom, or perhaps not independent of the kind, the circumstances, or the number of deviations. In some contexts, especially harsh punishments may not be *acceptable*, and hence certain allocations may be used to incentivize the agents in some states of the world, but not in others; also, depending on the states, the designer himself may be able to commit to certain outcomes of the mechanism, but not to others. As we will explain, whenever these considerations are present, the insights we receive from the classical literature that ignores such concerns for deviations are not applicable. We provide some examples:

(i) In a juridical context, for instance, prescribing punishments and rewards in response to 'deviant' behavior are often restricted by other constraints or desiderata, such as constitutional rights, higher level legislation, culture, or social norms.[3]

(ii) A central banker wants to allocate loans to commercial banks in a way that leads to the optimal level of financial stability for the economy. To do so, information about commercial banks' characteristics, such as current loan policies, is needed. This information is known to the commercial banks, but not to the central banker. But the central banker also wants to ensure that a minimal level of stability is reached even in the event that some commercial banks have incorrectly interpreted the current state of the system therefore leading to an incorrect report.

(iii) A competition authority trying to induce a certain market arrangement, which depends on information that is only available to the firms, may be subject to political constraints that limit its ability to commit to using certain punishments and rewards at certain states (see Ex. 1 below).

(iv) Furthermore, even if the designer manages to implement a given SCC with respect to a particular solution concept (say, Nash Equilibrium), he may still care that the outcomes associated with deviations are also *acceptable*, or very close the first-best 'target' allocation, if he is concerned for instance that the agents may make mistakes, or that they exhibit various forms of bounded rationality, or that their preferences are misspecified, and so on.[4]

[1] For instance, Nash (Maskin, 1999) and Subgame Perfect (Moore and Repullo, 1988), or more recently Rationalizable (Bergemann et al. (2011), Jain et al. (2022); Jain and Lombardi (2022)), Level-k (De Clippel et al., 2019), and Behavioral (De Clippel, 2014) Implementation. Maskin and Sjöström (2002) survey the early literature.

[2] Restrictions on the mechanisms have sometimes been imposed, for instance to avoid some unrealistic features of standard constructions in the literature (e.g., Jackson 1992), or to favor their economic interpretability (e.g., Ollár and Penta 2017, 2022, 2023), etc., but by and large the literature has not paid attention to the outcomes that a mechanism may induce, other than at the profiles that are consistent with the solution concept. Some exceptions are Bochet and Tumennasan (2022b,a), and the most closely related Eliaz (2002) and Shoukry (2019), which will be discussed extensively in the following.

[3] Juridical problems have been among the prime class of institutions about which implementation theory has been insightful. The recent literature on implementation with evidence, for instance (cf. Kartik and Tercieux (2012); Ben-Porath et al. (2019), etc.), is largely motivated by this kind of application, although it did not tackle the aspects that we will focus on, i.e. the designer's constraints on the outcomes induced by agents' deviations.

[4] For instance, in a famous example from Gneezy and Rustichini (2000) and popularized by Levitt and Dubner

To account for these considerations, we enrich the baseline framework by adding an *acceptability correspondence* that specifies, for each state of the world, the set of allocations that the designer wishes to ensure, if up to k agents deviate from the profiles that are consistent with the solution concept at that state. The resulting notion of *Safe Implementation* thus requires that, besides achieving implementation, a safe mechanism should also ensure that outcomes arising from up to k deviations are still acceptable to the designer. Besides the illustrative examples above, this notion provides a flexible framework to study a variety of robustness notions, related to a mechanism's safety and resilience properties, and it may also accommodate important and understudied problems within the implementation literature, such as the case of state-dependent feasible outcomes (see, e.g., Postlewaite and Wettstein 1989), limited commitment on the designer's part (see Example 1 below), a variety of robustness concerns, behavioral considerations, and others.

This modeling change, however, raises a number of challenges and conceptual innovations. In particular, the fact that both the SCC and the acceptability correspondence depend on the state of the world opens the door to a non-trivial interplay between the various elements of the model. This is due to the tension between the necessity to elicit the state of the world, the outcomes that need to be implemented, and the punishments that the designer can use to discipline agents' behavior, which are state-dependent themselves. Intuitively, if achieving standard (i.e., non-safe) implementation can be thought of as providing agents with the incentives to reveal the state, through a suitable scheme of punishments and rewards, Safe Implementation implies that the punishments that can be used are themselves restricted by the very information they are designed to extract. Hence, not only must agents be given the incentives to induce socially desirable allocations, but also to reveal which prizes and punishments can be used to achieve this task.

This interplay becomes apparent in the necessary and sufficient conditions that we provide for *Safe Nash Implementation*, i.e. when the underlying solution concept that describes agents' strategic interaction is taken to be Nash Equilibrium.[5] Our main necessary condition, which we call *Comonotonicity*, entails a joint restriction on the structure of the SCC and of the acceptability correspondence. For single-valued SCC (or Social Choice Functions, SCF), for instance, if Maskin Monotonicity (the famous necessary condition for Nash Implementation) requires that an allocation that is selected by the SCF at one state must also be selected at any other state in which it has (weakly) climbed up in all agents' rankings of the feasible alternatives, *Comonotonicity* strengthens the baseline notion in two ways: first, it states that for such an allocation to be selected by the SCF at the second state, it suffices that it climbs (weakly) up in everyone's ranking *only* compared to the alternatives that are acceptable at the first state; second, it requires the acceptability correspondence (not the SCF) to satisfy a form of monotonicity akin to Maskin's. As for sufficiency, our results show that *Comonotonicity* is almost sufficient as well, since it always ensures *Safe Nash Implementation* in combination with a generalization of Maskin's No-Veto condition that we call

(2006), in designing the Day Care fees, the Israeli public authority set up a system of fines for late retrievals that in the intention of the designer should have led parents to pick up their children on-time. But it turned out that a significant portion of parents deviated, finding it preferable to pay the fines and postpone the pick up time. This is an instance in which the outcome of a mechanism differed from the target it was designed to implement, due to deviations of a number of agents whose preferences were misspecified in the model used by the designer.

[5] While we do have results on Safe Implementation under general solution concepts, which will be discussed below, our main focus is on Nash Equilibrium, since it provides the classical workhorse for conceptual innovations within implementation theory. See, for instance, Kartik and Tercieux (2012) and Ben-Porath et al. (2019) for evidence-based implementation, Kartik et al. (2014) and Lombardi and Yoshihara (2020) for preferences for honesty, De Clippel (2014) for Behavioral Implementation, Hayashi and Lombardi (2017, 2019) for constrained implementation, etc.

Safe No-Veto, which is often automatically satisfied.[6] We note that both *Comonotonicity* and *Safe No-Veto* coincide with Maskin's conditions whenever the acceptability correspondence is vacuous (in the sense of admitting all outcomes at every state), in which case Safe Nash Implementation also coincides with (non-safe) Nash Implementation; but they are stronger in general. For the necessity part of our results, this is because the safety requirement that we impose does make implementation harder to obtain, and the conditions we provide directly reflect the extent to which this is the case.[7] Consider the following example:

Example 1 (Competition Policy with Non-Credible Punishments) Three firms, $1, 2$ and 3, are monopolists within their respective countries. While currently active only on their local markets, firms 1 and 2 could operate in any country. Firm 3 instead is a highly indebted company, who can only operate in its own country. A competition authority needs to choose between maintaining the status quo (allocation a), or changing the level of competition in the three markets by implementing alternatives b or c. In alternative b, all firms are active on all markets they can access, sharing each of them equally with the other firms with which they compete. Alternative c instead is the same as the status quo, except that the regulator lets firm 3 go bankrupt, splits 3's market equally between 1 and 2, but these firms must each pay half of the debt of the firm gone burst. For the sake of the example, these are the only feasible alternatives: $X = \{a, b, c\}$.

There are three states of the world, that reflect the state of the demand in market 3, which can be low (L), medium (M) or high (H). The true state is known to the firms but not to the designer. Firm 3's ranking is such that the status quo is always at the top. When the demand is low, then he prefers to be bailed out rather than face others' competition, but not when the demand is high or medium. Hence, its ranking is such that $a \succ b \succ c$, except if the demand is low, when it is $a \succ c \succ b$. As for the other firms, when demand in country 3 is medium, both firms 1 and 2 prefer to compete with each other in their local market, in order to access market 3 at no cost, but they would not be willing to enter it (even if splitting it in half) if they have to pay 3's debt. Hence, their ranking at this state is $b \succ a \succ c$. When the demand in country 3 is low, neither firm 1 nor 2 are willing to give up their monopolies in order to access the third market, and their ranking is $a \succ b \succ c$. When the demand is high, instead, both firms 1 and 2 prefer to absorb half of the third market and pay the debt of the bankrupt firm, over the status quo, over the fully competitive outcome in all countries. Their preferences therefore are $c \succ a \succ b$. (See Fig.1.)

Ideally, the competition authority would like to induce the competitive outcome, b, unless all firms prefer to maintain the status quo. Then, the SCF they wish to implement is such that $f(L) = a$ and $f(M) = f(H) = b$. Based on Maskin's results, absent safety concerns or restrictions on the implementing mechanism, it turns out that this SCF is Nash Implementable in this setting.

But now suppose that alternative c is not acceptable at the states where it is at the bottom for a majority of the firms, even as the outcome of a punishment designed to implement the SCF above. This may be because it would not be desirable for the designer to let firm 3 go bankrupt, or because it would not be politically credible to commit to enforcing such an outcome, if needed, in

[6] For our general results on SCC, we distinguish between a *weak* and a *strong* version of Comonotonicity. The two notions coincide for SCF. For SCC, the first notion is necessary, the second is for sufficiency.

[7] This result highlights an important difference between our approach and Eliaz's (2002). Namely that, unlike in our approach, the restrictions on the mechanism in Eliaz (2002) cannot be thought of as an *extra* desideratum on top of Nash implementation. In fact, implementation in the sense of Eliaz (2002) may obtain even if Nash Implementation is impossible. This is reflected in the necessary condition that he obtains, which unlike ours is not stronger than Maskin Monotonicity. This point will be further explained below.

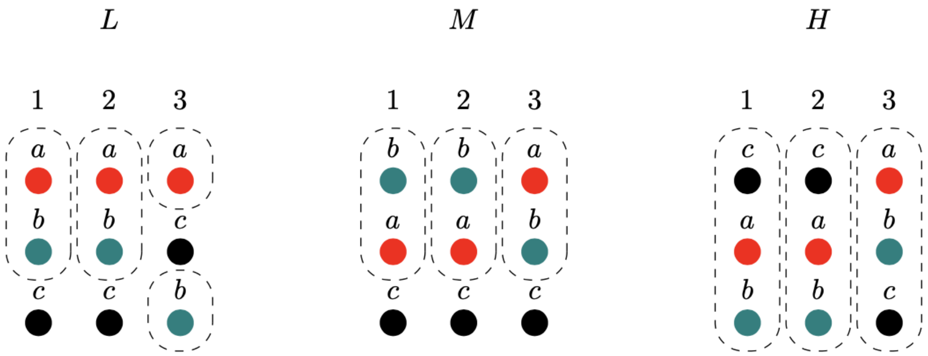

Figure 1: Firms 1, 2 and 3's preference orderings over the three alternatives, at the three states, L, M, and H. The acceptability correspondence, shown in dashed lines, is such that $A(L) = A(M) = \{a, b\}$, $A(H) = \{a, b, c\}$. In this setting, the SCF such that $f(L) = a$ and $f(M) = f(H) = b$ is Nash Implementable, but not Safely so, with respect to acceptability correspondence A.

response to someone's deviation (for instance, the three firms can be from three different European countries, and it may not be credible that the competition authority would get the political support to let country's 3 firm go burst, if needed, at a state when it is the worst outcome for the majority). That is, suppose that outcome c does not belong to the acceptability correspondence at states L and M. Then, it turns out that the SCF above cannot be Safely Implemented in this case. Thus, if the designer is subject to such political constraints, which make outcome c not credible at some states, then the insights based on the classical results are misleading.

Specifically, our results imply that in order to fulfill the Safety requirement, the designer in this case must settle for the status quo also at state H, thereby implementing a SCF that induces the competitive outcome less often. The intuition is that if b and not a has to be selected at state H (as entailed by SCF f above), in order to avoid the existence of a Nash equilibrium at H in which firms collude so as to induce the non-competitive outcome, the designer must rely on outcome c as a deterrent, since at such a state all agents prefer a over b. But if this were allowed, then c could emerge as the outcome of a deviation from an equilibrium at state L, where it is not acceptable. As a consequence, c cannot be used to discipline behavior at state H either, and hence only a SCF that chooses the same outcome at both L and H can be implemented. □

After providing the general necessary and sufficient conditions for Safe Nash Implementation, we move on to consider special cases of interest, in which we provide both positive and negative results. For instance, in economies that satisfy a standard single-crossing condition, we show that any SCF can be Safely Nash Implemented, whenever the acceptability correspondence at every state includes an arbitrarily small neighborhood of the allocation prescribed by the SCF. This means that, in these settings, any SCF can be implemented in the *Almost Perfectly Safe* sense, i.e. ensuring that the allocation remains arbitrarily close to the desired one even if up to k agents deviate from the equilibrium profiles, for any $k < \frac{n}{2}$ (where n is the number of agents in the economy). The intuition for this result is that, in environments with a continuum and convex outcome space, and if preferences are continuous and satisfy standard single-crossing properties, incentives can effectively be provided with small deviations from the allocations that the designer wishes to implement. This insight is clearly in stark contrast with Example 1 above, which obviously features indivisibilities. Indeed, it is generally the case that safety concerns are harder to accommodate when indivisibilities are present, or in the absence of transfers. Nonetheless, as we show in Section 6, positive results can also be obtained in important economic settings with indivisibilities. Specifically, in assignment problems of one unit of an indivisible good, we show that the efficient allocation can always be Safely Nash Implemented, whenever there is some *null allocation* that is included in the acceptability correspondence at all states.

The results above show that there are interesting and important economic environments in which Safety concerns can be accommodated at minimal or no cost. But Safe Implementation also has its limits: as we further show, seemingly plausible safety requirements can never be implemented, regardless of the underlying solution concept (be it Nash Equilibrium or not), when preferences are 'rich' or when the SCF is surjective on the space of feasible allocations. Thus, safety requirements are demanding in general, and there are serious limits to their implementability. Nonetheless, economically important settings exist in which they can be guaranteed under standard and generally weak conditions.

Our approach also provides a methodological contribution to the literature in mechanism design and implementation that aims to incorporate behavioral notions or robustness to mistakes in play (see, e.g., Eliaz (2002); Renou and Schlag (2011); Tumennasan (2013); De Clippel (2014), De Clippel et al. (2019), Crawford (2021), etc.). These objectives have typically been addressed via suitable modifications of the solution concepts that underlie the notion of implementation. This has led to a variety of notions, each tailored to a specific behavioral departure from the standard model. But misspecification of the behavioral model itself is typically not allowed. In contrast, by incorporating robustness concerns as restrictions on the outcomes that may ensue as the result of players' *deviations*, our approach complements the literature by not forcing the chosen model to be perfectly specified. While we have mainly focused on Nash equilibrium, our framework is flexible and can be applied to *any* solution concept, be it classical or behavioral. In that sense, it can advance the agenda of this literature, not only by favoring its integration with standard notions (such as Nash implementation), but also by providing a 'detail free' way of accounting for the possibility of behavioral deviations, without necessarily ascribing to a particular theory thereof.

2 Model

Preliminaries: We consider environments with complete information, with a finite set of agents, $N = \{1, ..., n\}$, and an outcome space X. Each agent i has a bounded utility $u_i : X \times \Theta \to \mathbb{R}$, where Θ is the set of states of nature, with typical element $\theta \in \Theta$, which we assume is commonly known by the agents unknown to the designer. We let $\mathcal{E} = \left\langle N, \Theta, X, (u_i)_{i\in N} \right\rangle$ denote the environment from the viewpoint of the designer, and for any $\theta \in \Theta$, we let $\mathcal{E}(\theta) := \left\langle N, X, (u_i(\cdot, \theta))_{i\in N} \right\rangle$ denote the environment in which agents commonly know that preferences are $(u_i(\cdot, \theta))_{i\in N}$. Finally, for any $i \in N$, $\theta \in \Theta$ and $x \in X$, we let $L_i(x, \theta) := \{y \in X : u_i(y, \theta) \leq u_i(x, \theta)\}$ denote agent i's lower contour set of outcome x in state θ.

A social planner aims to choose an outcome (or a set of outcomes), as a function of the state of nature. These objectives are represented by a *social choice correspondence* (SCC), $F : \Theta \to 2^X \setminus \emptyset$, that assigns a (non-empty) set of outcomes to each state of nature. The special case in which $F(\theta)$ is a singleton for every θ is referred to as *social choice function* (SCF), and denoted by

$f : \Theta \to X$. States of nature are known to the agents but not to the designer. Thus, in a standard implementation problem, the designer's problem is to design a mechanism with the objective that, letting players interact, given their knowledge of the state of the world, their behavior in the mechanism induces outcomes that are included in the SCC correspondence for any state.

Formally, a *mechanism* is a tuple $\mathcal{M} = \langle (M_i)_{i\in N}, g\rangle$, where for each $i \in N$, M_i denotes the set of messages of agent i, and $g : M \to X$ is an outcome function that assigns one allocation to each message profile, where we let $M = \times_{i\in N} M_i$ and $M_{-i} = \times_{j\neq i} M_j$. Similarly, for subsets of players $D \subset N$, we let M_D and M_{-D} denote, respectively, the set of message profiles of all agents that are inside and outside the set D. For each $\theta \in \Theta$, any mechanism $\mathcal{M} = \langle (M_i)_{i\in N}, g\rangle$ induces a complete information game $G^{\mathcal{M}}(\theta) := \langle N, (M_i, U_i^{\theta})_{i\in N}\rangle$, where M_i is the set of strategies of player i, and payoff functions are such that $U_i^{\theta}(m) = u_i(g(m), \theta)$ for all $i \in N$ and $m \in M$. Agents' behavior is described by a *solution concept*, $\mathcal{C}$, which for any given mechanism $\mathcal{M}$ induces a correspondence $\mathcal{C}^{\mathcal{M}} : \Theta \to 2^M$ that assigns a (possibly empty) set of message profiles to every state of the world. For any mechanism $\mathcal{M} = \langle (M_i)_{i\in N}, g\rangle$ and state $\theta \in \Theta$, we let $g(\mathcal{C}^{\mathcal{M}}(\theta)) := \{x \in X : \exists m \in \mathcal{C}^{\mathcal{M}}(\theta) : g(m) = x\}$ denote the set of outcomes that are induced by action profiles that are consistent with the solution concept $\mathcal{C}$, at the state of the world θ. Full (strong) implementation is defined as follows:

Definition 1 (Implementation) *A SCC is (fully) $\mathcal{C}$-implementable (or, it is fully implementable with respect to solution concept $\mathcal{C}$), if there exists some mechanism $\mathcal{M}$ s.t. (i) $\mathcal{C}^{\mathcal{M}}(\theta) \neq \emptyset$, and (ii) $g(\mathcal{C}^{\mathcal{M}}(\theta)) = F(\theta)$ for all $\theta \in \Theta$.*

For instance, if $\mathcal{C}$ is such that $\mathcal{C}^{\mathcal{M}}(\theta)$ denotes the set of Nash Equilibria of $G^{\mathcal{M}}(\theta)$ (i.e., $\mathcal{C}^{\mathcal{M}}(\theta) := \{m^* \in M : \forall i \in N, U_i^{\theta}(m^*) \geq U_i^{\theta}(m_i, m^*_{-i})\}$), then the standard notion of *Nash Implementation* (Maskin, 1999) obtains.

Safe Implementation: Next we introduce the elements of the model that are needed for the social choice correspondence to be *safely* implemented. As we discussed in the introduction, the idea is that the designer not only wishes to attain $\mathcal{C}$-implementation, but also ensure that the implementing mechanism has the property that, should a number of agents deviate (perhaps due to irrationality, a mistake, or because the planner's model of their preferences or of their behavior is misspecified), the mechanism still induces outcomes that the designer regards as *acceptable*. Like the 'target' allocations in the SCC, however, also what is regarded as *acceptable* may depend on the state. This is modelled by an *acceptability correspondence*, $A : \Theta \to 2^X \setminus \emptyset$, where $A(\theta)$ denotes the set of outcomes that the social planner regards as acceptable at state θ. A natural requirement – which, in fact, would follow immediately as a necessary condition from Def. 2 below, and which therefore we maintain throughout – is that $F(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$.

Example 2 ***(Some Examples and Special Cases)***

1. *Minimal Safety Guarantees:* In some settings, it may be natural for the social planner to impose a minimal safety guarantee in the sense that, in the result of deviations from equilibrium, it ensures that no agent receives their least preferred outcome. We say that an acceptability correspondence $A : \Theta \to 2^X \setminus \emptyset$ is *minimally safeguarding* if, for all $\theta \in \Theta$,

$$A(\theta) = X \setminus \left\{ x \in X : \exists j \in N \quad s.t. \quad x \in \operatorname*{argmin}_{x\in X} u_j(x,\theta) \setminus \operatorname*{argmax}_{x\in X} u_j(x,\theta) \right\} \tag{1}$$

2. *Planner's Welfare Guarantees:* The acceptability correspondence may explicitly represent the concerns of a social planner under second best considerations. For instance, if the planner has state-dependent preferences over allocations, $W : X \times \Theta \to \mathbb{R}$, then it is natural to think about the SCC as the set of *optimal* outcomes at every state (i.e., $F(\theta) = \arg\max_{x \in X} W(x, \theta)$ for all θ), and to consider *acceptable* allocations that ensure that the planner attains at least a certain (possibly state-dependent) reservation value $\bar{w}(\theta)$. In this case, the acceptability correspondence is defined such that, for all $\theta \in \Theta$, $A(\theta) = \{x \in X : W(x, \theta) \geq \bar{w}(\theta)\}$. For instance, the planner may only be willing to sacrifice a fraction $\alpha \in (0, 1)$ of the optimal welfare when he punishes deviations, and hence set $\bar{w}(\theta) = (1 - \alpha) \max_{x \in X} W(x, \theta)$. In this case, W may represent different welfare functions, such as a generalized utilitarian (i.e., $W(x, \theta) = \sum_{i \in N} \lambda_i u_i(x, \theta)$ for some $(\lambda_i)_{i \in N} \in \mathbb{R}^n_+ \setminus \{0\}$), Rawlsian (i.e., $W(x, \theta) = \min_{j \in N} u_j(x, \theta)$), or other social welfare criteria.

3. *Perfect Safety:* Another interesting special case is when $A(\theta) = F(\theta)$ for all $\theta \in \Theta$. This is in a sense the most demanding notion of safety, in that it requires that also the deviations do not induce outcomes inconsistent with the SCC.

4. *ϵ-Perfect Safety:* When X is a metric space, one reasonable restriction is that the acceptable allocations are within a given distance from the choices in the SCC or SCF. For instance, one could define $A(\theta) = \mathcal{N}_\epsilon(f(\theta))$ for all $\theta \in \Theta$, where $\mathcal{N}_\epsilon$ is an epsilon neighbourhood with respect to the metric on X. In this sense, the acceptable allocations would be close to the 'optimal' ones in the literal sense.

5. *Limited Commitment Interpretation:* In the previous examples the acceptability correspondence is derived from welfare considerations that the planner may have in mind. More broadly, however, $A(\cdot)$ may represent other constraints that the planner faces in designing the mechanism, and particularly the outcomes after players' deviations, which may serve as punishments to provide agents with the incentives to induce socially desirable allocations. In designing such punishments, however, the designer may be constrained in what he can commit to, and for instance mechanisms that prescribe especially harsh punishments may not be credible at certain states after a small number of deviations. From that viewpoint, for each θ, $A(\theta)$ can be taken as a primitive that encompasses the set of outcomes that the planner can credibly commit to using as punishments at that state.

6. *State-Dependent Feasible Allocations:* Our framework can also be used to accommodate the case in which the very set of feasible allocations is state-dependent, and the outcomes of a mechanism are required to be feasible not only at equilibrium, but also after deviations. This problem has been studied, for instance, by Postlewaite and Wettstein (1989) in the context of Walrasian Implementation, in a setting in which the state of the world includes not only agents' preferences but also their initial endowments, and hence the set of feasible allocations is unknown to the designer. Within this setting, Postlewaite and Wettstein (1989) provide a mechanism that Nash-implements the Walrasian correspondence – as well as achieve other desiderata, such as a continuous outcome function – under state-dependent feasibility restrictions. Obviously, the case of state-dependent feasible allocations is relevant in a variety of settings, other than Walrasian implementation, but it has been surprisingly

neglected. It can be accommodated within our framework simply by reinterpreting each $A(\theta)$ as the set of allocations that are feasible at state θ. Our necessity results (Theorems 1 and 2) therefore directly imply necessary conditions for implementation with state-dependent feasible allocations, for general SCC, thereby filling an important gap in the literature.

Next, let $k \in \{1, ..., n\}$ denote the *safety level* that the designer wishes to impose. That is, the maximum number of deviations from the solutions $m^* \in \mathcal{C}^{\mathcal{M}}(\theta)$ that the designer wants to ensure they induce outcomes in $A(\theta)$, for all θ. If $k = n$, then the safety level is such that the mechanism is never allowed to select an allocation outside of $A(\theta)$ in any state of the world. This is the relevant case, for instance, if one reinterprets $A(\theta)$ as the (state-dependent) set of feasible allocations, as for instance in Postlewaite and Wettstein (1989) that we just discussed (point 6 in Ex. 2). The other especially relevant case is when $k = 1$. In this case, like baseline Nash Implementation, (A, k)-Safe Implementation is only concerned with *unilateral* deviations, but it requires that they are not only *unprofitable* for the agents, but also *acceptable* to the designer.

For any $k \in \{1, ..., n\}$ let N_k denote the set of all subsets of N with k elements (that is, $N_k := \{C \in 2^N : |C| = k\}$), and further define a distance function $d_N(m, m') := |\{i \in N : m_i \neq m'_i\}|$ and a neighbourhood $B_k(m) := \{m' \in M : d_N(m, m') \leq k\}$, which consists of the set of message profiles m' that differ from m for at most k messages. Also, we say that $A^* : \Theta \to 2^X \setminus \emptyset$ is a *sub-correspondence* of $A : \Theta \to 2^X \setminus \emptyset$ if it is such that $A^*(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$. With this, (A, k)-Safe Implementation is defined as follows:[8]

Definition 2 ((A, k) Safe Implementation) *Fix a solution concept $\mathcal{C}$, $k \in \{1, ..., n\}$, a SCC $F : \theta \to 2^X \setminus \emptyset$, and let $A : \Theta \to 2^X \setminus \emptyset$ denote an* acceptability correspondence*, such that $F(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$. We say that F is (A, k)-Safe $\mathcal{C}$-implementable if there exists a mechanism $\mathcal{M} = ((M_i)_{i \in N}, g)$ such that: (i) F is $\mathcal{C}$-Implemented by $\mathcal{M}$, and (ii) for all $\theta \in \Theta$, $m^* \in \mathcal{C}(\theta)$, and for all $m' \in B_k(m^*)$, $g(m') \in A(\theta)$.*

If, furthermore, the acceptability correspondence, A, admits no sub-correspondence A^* for which (A^*, k)-Safe $\mathcal{C}$-Implementation is possible, then we say that A is maximally safe.

First note that, for any $\mathcal{C}$, this notion generalizes the standard notion of (non-safe) implementation of Def. 1, which obtains as the special case in which condition (ii) in this definition is moot, which is the case for any k if $A(\theta) = X$ for all θ. As we will discuss, Def. 2 also generalizes existing notions in the literature, such as *outcome-robust implementation* of SCF in Shoukry (2019), and *Fault Tolerant Implementation* (Eliaz, 2002), that share a similar motivation to ours.[9]

Second, for any k, if a SCC is (A, k)-Safe Implementable, then it is $(\hat{A}, k)$-Safe Implementable for any 'more permissive' correspondence, $\hat{A} : \Theta \to 2^X \setminus \emptyset$, such that $A(\theta) \subseteq \hat{A}(\theta)$ for all $\theta \in \Theta$. This observation motivates the notion of **Maximally Safe** acceptability correspondence in Def. 2: if a SCC is (A, k)-Safe $\mathcal{C}$-Implementation, but not with respect to any sub-correspondence of A, then it means that A reflects the most demanding acceptability correspondence that the designer could impose, while still retaining Safety.

[8] Most of our analysis will focus on the case in which $\mathcal{C}$ is Nash Equilibrium. Nonetheless, this general definition is useful to clarify the connections with the related literature, and to provide some general results.

[9] Both of these papers will be discussed extensively. Shoukry (2014) considers restrictions similar to minimal safety guarantees, but as we discuss in Section 7, the main notion of implementation in that paper is very different.

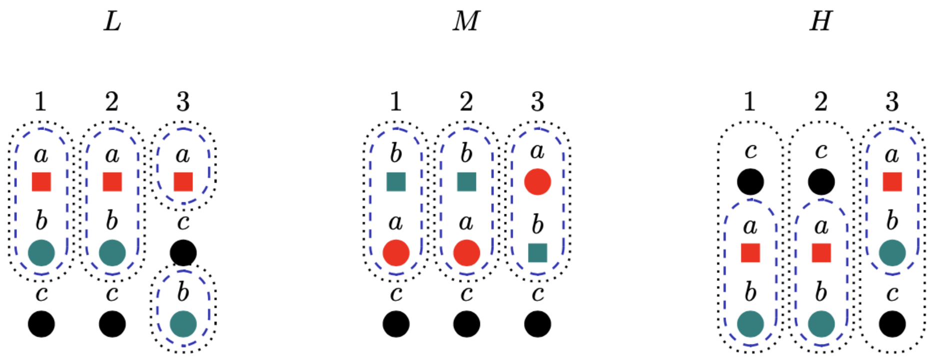

Figure 2: Firms 1, 2 and 3's preference orderings over the three alternatives, at the three states, L, M, and H. For each state, the allocation chosen by SCF f^* in Ex. 3 is indicated by a square. The acceptability correspondence A from Ex.1 is shown by the dotted lines, and is not maximally safe for this SCF. Acceptability correspondence A^* in Ex. 3 is maximally safe, and is represented by the dashed lines in the figure.

Example 3 Consider again the environment in Ex.1: it will follow from our results that a SCF such that $f^*(L) = f^*(H) = a$ and $f^*(M) = b$ is Safe Implementable (letting the solution concept, $\mathcal{C}$, be Nash equilibrium) with respect to the A correspondence in Ex.1 (see Fig.2). That acceptability correspondence, however, is not *maximally safe* for such a SCF, because it can be shown that the same SCF can also be Safe Implemented with respect to a sub-correspondence of A that rules out outcome c also at state H. Formally, $A^* : \Theta \to 2^X \setminus \emptyset$ s.t. $A^*(\theta) = \{a, b\}$ for all θ. □

With this in mind, it should also be clear that the case $A(\theta) = F(\theta)$ for all $\theta \in \Theta$ (case 3 in Ex.2) is the most demanding case (albeit not necessarily possible, depending on F), and will be referred to as **Perfectly Safe Implementation**.[10] We will instead use the term **Almost Perfectly Safe Implementation** to refer to the case in which, *for all* $\epsilon > 0$, Safe Implementation can be obtained with respect to an ϵ-Perfectly Safe acceptability correspondence (case 4 in Ex.2).

Third, if the solution concept is held fixed across k (for instance, if $\mathcal{C}$ is taken to be Nash Equilibrium, as we will do in the following sections), then for any acceptability correspondence $A : \Theta \to 2^X \setminus \emptyset$, a SCC is (A, k)-Safe Implementable only if it is (A, k')-Safe Implementable for all $k' \leq k$. That is, if $\mathcal{C}$ is constant, then increasing the k parameter does make the safety requirement more demanding. This is not necessarily true if instead the solution concept depends on k, as it is the case for instance with the notion of *Fault Tolerant Implementation* (FTI), in which implementation may fail for some k, and be possible for some $k' > k$.[11]

[10] For the case of SCF, Shoukry (2019)'s *outcome-robust implementation* corresponds to this case, with $\mathcal{C}$ equal to Nash Equilibrium, but allowing transfers and assuming that players have preferences for truthtelling. For the case of non-single valued SCC, his notion is more restrictive than Perfectly Safe Implementation, since not only it requires that the outcome stays within the SCC, but that it doesn't change at all.

[11] Fault Tolerant Implementation (Eliaz, 2002) obtains as the special case of Def.2, letting the acceptability correspondence be such that $A(\theta) = F(\theta)$ for all $\theta \in \Theta$, and taking as solution concept the so called *k-Fault Tolerant Nash Equilibrium (k-FTNE)*. The reason why, under this notion, increasing k does not necessarily tighten the implementation requirement. That is that k-FTNE depends on k in such a way that, if $\mathcal{M}$ is such that $\mathcal{C}_k^{\mathcal{M}}(\theta) \neq \emptyset \neq \mathcal{C}_{k'}^{\mathcal{M}}(\theta)$, and $k' < k$, then it may be that $\mathcal{C}_k^{\mathcal{M}}(\theta) \subseteq \mathcal{C}_{k'}^{\mathcal{M}}(\theta)$. The monotonicity in k that holds when $\mathcal{C}$ is held constant (as is the case for the main focus of our paper, in which $\mathcal{C}$ is standard Nash Equilibrium) is thus not guaranteed for FTI.

3 Safe Nash Implementation

For the time being, we will take Nash Equilibrium to be the underlying solution concept, and hence for any mechanism $\mathcal{M}$, the correspondence $\mathcal{C}^{\mathcal{M}} : \Theta \to 2^M$ in Definition 2 coincides with the Nash Equilibrium correspondence. This is what we refer to as **Safe Nash Implementation**. (We will return to general solution concepts in Section 6.3). Also, as in Def. 2, if the acceptability correspondence A is such that Safe Nash Implementation is impossible for all sub-correspondences, then we say that A is **maximally safe**. For ease of reference, we reproduce here these definitions:

Definition 3 (Safe Nash Implementation) *A SCC is (A, k)-Safely Nash Implementable if it is Nash Implemented by a mechanism $\mathcal{M} = ((M_i)_{i\in N}, g)$ such that for all states θ, for all Nash equilibria m^* of $G^{\mathcal{M}}(\theta)$, and for all $m \in B_k(m^*)$, $g(m) \in A(\theta)$.*

If, furthermore, the acceptability correspondence, A, admits no sub-correspondence A^* for which (A^*, k)-Safe Nash Implementation is possible, then we say that A is maximally safe*.*

The natural benchmark is obviously Nash Implementation (Maskin, 1999), which obtains as a special case of (A, k)-Safe Nash Implementation when the extra safety requirement is moot (which is the case, as we mentioned, if $A(\theta) = X$ for all $\theta \in \Theta$). Also, it is straightforward to check that the following hold: (i) if a SCC is (A, k)-Safe Nash Implementable, then it is $(\hat{A}, k)$-Safe Nash Implementable for all $\hat{A} : \Theta \to 2^X \setminus \emptyset$ s.t. $A(\theta) \subseteq \hat{A}(\theta)$ for all $\theta \in \Theta$ – that is, making the acceptability correspondence more permissive makes implementation easier to achieve; (ii) since the solution concept does not depend on k, if a SCC is (A, k)-Safe Nash Implementable, then it is (A, k')-Safe Nash Implementation for all $k' \leq k$ – that is, increasing the number of deviations the mechanism must be resilient to makes implementation harder.

As it is well known, Maskin (1999) showed that the following condition is necessary for (non-safe) Nash Implementation:

Definition 4 (Maskin Monotonicity) *A SCC is (Maskin) monotonic if for any θ, θ', if $x \in F(\theta)$ is such that $L_i(x, \theta) \subseteq L_i(x, \theta')$ for every $i \in N$, then $x \in F(\theta')$.*

Maskin (1999) also showed that, together with the following 'no veto condition', monotonicity is also sufficient for Nash Implementation, whenever $n \geq 3$:

Definition 5 (Maskin No Veto) *A SCF satisfies the 'no veto property' if whenever θ is such that there exist $x \in X$ and $i \in N$ s.t. $x \in \arg\max_{y\in X} u_j(y, \theta)$ for all $j \neq i$, then $x \in F(\theta)$.*

Obviously, Def. 5 has no bite if preferences rule out 'almost unanimity', as is the case in *economic environments*, where agents have strictly opposing interests (e.g., Mirrlees (1976), Spence (1980), Arya et al. (2000), Kartik and Tercieux (2012), etc.). Formally:

Definition 6 (Economic Environments (cf., Kartik and Tercieux (2012)) *An environment is* economic *if for all $\theta \in \Theta$ and $x \in X$, $|\{i \in N : x \in argmax_{y\in X} u_i(y, \theta)\}| < n - 1$.*

Thus, in economic environments, monotonicity is both necessary and sufficient for (non-safe) Nash implementation. In the next two sections we provide necessary and sufficient conditions for

Safe Nash Implementation. Since Nash Implementation is a special case of Safe Nash Implementation, the necessary conditions for Safe Nash-Implementation will have to be a generalization of Definition 4. Our sufficient conditions will also be a generalization of Maskin's, and we will show that they coincide with the necessary conditions in environments that satisfy an 'economic condition' analogous to the one above, or if the designer is allowed to adopt stochastic mechanisms.

4 Necessity

We introduce next a generalization of (Maskin) Monotonicity, which will be shown to be necessary for (A, k)-Safe Nash Implementation:

Definition 7 (Weak Comonotonicity) *A SCC, $F : \Theta \to 2^X \setminus \emptyset$, and an acceptability correspondence, $A : \Theta \to 2^X \setminus \emptyset$, are* weakly comonotonic *if they satisfy the following conditions:*

1. [A-Constrained Monotonicity of F] *If $\theta, \theta' \in \Theta$ and $x \in F(\theta)$ are such that $L_i(x, \theta) \cap A(\theta) \subseteq L_i(x, \theta') \cap A(\theta)$ for all $i \in N$, then $x \in F(\theta')$.*

2. [weakly F-Constrained Monotonicity of A] *If $\theta, \theta' \in \Theta$ are such that, $\forall x \in F(\theta)$, $L_i(x, \theta) \cap A(\theta) \subseteq L_i(x, \theta') \cap A(\theta)$ for all $i \in N$, then $A(\theta) \subseteq A(\theta')$.*

To understand this condition, first note that weak Comonotonicity implies (Maskin) Monotonicity: If $\theta, \theta' \in \Theta$ are such that $L_i(x, \theta) \subseteq L_i(x, \theta')$, and $x \in F(\theta)$, then the condition in part 1 of Def. 7 is satisfied for any A, and hence $x \in F(\theta')$, as requested by Def. 4.

Second, if $A(\theta) = X$ for every θ – i.e., if the safety requirement is vacuous, and Safe Nash Implementation coincides with Nash Implementation – then part 2 in Def. 7 holds vacuously, letting $A = X$, and part 1 coincides with (Maskin) Monotonicity. But if the A-correspondence entails non-trivial acceptability restrictions, then part 1 of Def. 7 restricts the SCC more than (Maskin) Monotonicity does. For a SCF, for instance, this condition requires that $f(\theta) = f(\theta')$ whenever $L_i(f(\theta), \theta) \cap A(\theta) \subseteq L_i(f(\theta), \theta') \cap A(\theta)$, which may be the case even if $L_i(f(\theta), \theta) \nsubseteq L_i(f(\theta), \theta')$. In the latter case, (Maskin) Monotonicity alone would leave the SCF free to set $f(\theta') \neq f(\theta)$, but weak Comonotonicity would not (see Ex. 1 in the Introduction). Thus, when the acceptability correspondence is non-trivial, weak Comonotonicity forces the SCF to be relatively more constant than Maskin's monotonicity would, and more so as the acceptability correspondence gets less permissive. More broadly, note that part 1 of Def. 7 gets less restrictive as the admissibility correspondence gets more inclusive: if A satisfies part 1 of Def. 7, and $\hat{A}$ is such that $A(\theta) \subseteq \hat{A}(\theta)$ for all $\theta \in \Theta$, then also $\hat{A}$ satisfies it.

The second part of Def. 7 states a monotonicity property of the acceptability correspondence, akin to Maskin's monotonicity for SCC, which imposes a lower bound on its inclusivity. Its mechanics is perhaps easier to grasp by looking at the (equivalent) contrapositive statement of that condition. Namely, if some allocation is acceptable according to the A correspondence at state θ but not at state θ', then there must exist a 'target' allocation $x \in F(\theta)$ that, going from state θ to θ', has moved down in the ranking of the allocations within $A(\theta)$ for at least one of the agents. Note that, in this case, the bite of the condition depends on the SCC: the more inclusive the SCC, the less stringent part 2 of Def. 7. This suggests, for instance, that compared with the case of SCF, this condition leaves more freedom for the set of acceptable allocations to vary with the state when the designer aims to implement a (non single-valued) SCC.

Finally, we note that weak Comonotonicity is no harder to check than Maskin monotonicity (except, of course, that one also needs to check for the A correspondence, besides the F).[12]

We can now turn to our main results on necessity. As discussed in Section 2, Safe implementation becomes more restrictive as the A correspondence gets finer. Hence, as far as necessary conditions are concerned, it is natural to start with the case when the acceptability correspondence is *Maximally Safe*, which puts the most stringent constraints on safe implementation (if a SCC is (maximally) safe implementable with respect to A, then it would also be Safe-Implementable with respect to any 'coarser' acceptability correspondence, A^*, such that $A(\theta) \subseteq A^*(\theta)$ for all θ). We show next that weak Comonotonicity is necessary for *maximally* safe Nash implementation:

Theorem 1 (Necessity) *A SCC, $F : \Theta \to 2^X \setminus \emptyset$, is maximally (A, k)-Safe Nash Implementable* only if (F, A) *are weakly Comonotonic.*

To understand the intuition behind this result, note that if the SCC is (A, k)-Safe Nash Implementable, and A is maximally safe, then for each $\theta \in \Theta$, the set $A(\theta)$ comprises *all* the outcomes that the designer can use to deter agents' deviations, and no more than those. Thus, from the viewpoint of providing agents with the right incentives within the mechanism, at any given state θ, it is only agents' preferences over the set $A(\theta)$ that matter. So, if going from one state θ to another θ', one of the 'target' allocations x climbs (weakly) up in everyone's ranking *within the restricted set $A(\theta)$ of acceptable allocations* (not over all of X, as in (Maskin) Monotonicity), and if – by the Nash implementation requirement – x must be a Nash equilibrium outcome at state θ for some mechanism, then it would also have to be a Nash equilibrium outcome at state θ'. But then x should be within the SCC at both states, otherwise Nash implementation would not obtain. This explains the necessity of part 1 of Def. 7.

To understand part 2, if going from state θ to θ' we have that in fact *all* the allocations in $F(\theta)$ (weakly) 'climb up' in everyone's ranking within the $A(\theta)$ set, then *all* such allocations would be Nash Equilibrium outcomes at both states θ and θ', and would each be induced by some Nash equilibrium profile m^* in some mechanism. But then, in such a mechanism, the set of outcomes that are within k deviations from such m^* at state θ, would also be within k-deviations from a Nash equilibrium at state θ', and thus they must also be acceptable at that state, if Safe Implementation is achieved. It follows that $A(\theta')$ must contain at least all of the outcomes that are within k deviations from Nash equilibria at θ, and hence in $A(\theta)$.

As we discussed, moving to the case of non-maximally safe acceptability correspondences, Safe Nash implementation gets less demanding, and hence also the necessary conditions are expected to be weaker. Nonetheless, it is easy to see from the argument above that, if A is *not* maximally safe, then the first part of Def. 7 is still necessary. The second part, however, need not hold:

Example 4 Consider again the environment in Example 3 (see Fig.2). As discussed, the SCF f^* from that example is safe implementable with respect to both correspondences A and A^*, but only the latter is *maximally safe* with respect to f^* (A cannot be, since A^* is a sub-correspondence of A). It is easy to check that, as it follows from Theorem 1, A^* satisfies both conditions in Def.

[12] Arguably, this was not the case for a version we had in an earlier draft of the paper. We would like to stress that the issue is corrected in this formulation.

7, and hence that it is (weakly) comonotonic with respect f^*. In contrast, the A correspondence only satisfies part 1 of Def. 7 (as implied by Proposition 1), but not part 2: moving from state $\theta = H$ to $\theta' = L$, allocation $a = f^*(H)$ moves (weakly) up in everyone's ranking within the set $A(H) = \{a, b, c\}$. Yet, $A(H) \not\subseteq A(L)$. This is obviously not the case for the A^* correspondence, since $A^*(H) = A^*(L) = \{a, b\}$. $\square$

Proposition 1 (Non-maximally safe implementation: necessity) *A SCC, $F : \Theta \to 2^X \setminus \emptyset$, is (non-maximally) (A, k)-Safe Nash Implementable* only if (F, A) *satisfy part 1 of Def.* 7 *(that is, A-constrained Monotonicity of F).*

The necessity results above formalize a trade-off between the restrictiveness of the acceptability correspondence and the way in which the SCC correspondence varies with θ. This is easier to see considering the case of a SCF. Suppose that the designer starts with a (Maskin) Monotonic SCF (as discussed, this is the minimal necessary condition, and it coincides with weak Comonotonicity if the acceptability restriction is vacuous). Then, among the $A^* : \Theta \to 2^X \setminus \emptyset$ correspondences that satisfy parts 1 and 2 of Def.7, those (if they exist) that are minimal with respect to set inclusion at every state, identify the most demanding acceptability requirements that the designer can impose, if he wishes to achieve Safe Nash Implementation. If, however, the safety desiderata are more stringent than this (i.e., if no such $\subseteq$-minimal A^* is a sub-correspondence of the acceptability correspondence that the designer wishes to impose), then Safe Nash Implementation necessarily forces the SCF to be more constant than what is implied by (Maskin) Monotonicity (Ex.1 in the Introduction provides an instance of this. To see it further, it can be shown that if the acceptability requirement in our example from Fig. 2 were further shrunk, imposing a sub-correspondence of A^*, then only constant SCFs could be Safe-Implemented). Indeed, this is consistent with the intuition that the safety requirement makes implementation harder: Safe Implementation entails stronger necessary conditions than Nash Implementation.[13]

Theorem 1 also has the following direct and important implication:

Corollary 1 (Impossibility of Perfectly Safe Implementation of SCF) *For any $k \geq 1$, if $f : \Theta \to X$ and $A : \Theta \to 2^X \setminus \emptyset$ is s.t. $A(\theta) = \{f(\theta)\}$ for some θ, then f is (A, k)-Safely Nash Implementable only if f is constant. It follows that only constant SCFs can be Perfectly Safely Nash-Implemented.*

This result follows directly from part 1 of Def. 7: if $A(\theta) = \{f(\theta)\}$, then $L_i(f(\theta), \theta) \cap A(\theta) = \{f(\theta)\} \subseteq L_i(f(\theta), \theta')$ for any θ', and the necessity of Comonotonicity implies that f is (A, k)-Safely Nash Implementable only if $x = f(\theta')$ for all θ'.

Corollary 1 is especially relevant to understand the connection with the related notions put forward by Eliaz (2002) and Shoukry (2019), which for SCFs are a special case of Def. 2 in which the acceptability correspondence is set to be the most demanding, in that it requires *Perfect Safety* (cf. point 3 in Ex. 2). More specifically, Corollary 1 suggests a certain trade-off between the restrictiveness of the acceptability correspondence and the *solution concept* underlying the notion of implementation. In Eliaz (2002), for instance, positive results for non-constant SCFs are made possible by the weakening of the implementation requirement due to the adoption of a refinement

[13] As already mentioned, this is not the case for notions in which the solution concept varies with k, as in Eliaz (2002). This point is further discussed below (see also footnote 11).

of Nash Equilibrium: since k-FTNE refines Nash Equilibrium (and more so, as k increases), it makes it easier to avoid 'bad' equilibria.[14] Shoukry (2019), instead, maintains both the Perfect Safety requirement and Nash Equilibrium as a solution concept, and in order to recover possibility results for SCFs, he allows for transfers and a preference for the truth.[15]

Despite this impossibility of *Perfectly Safe* Nash Implementation, however, we will show that in an important class of environments it is possible to get arbitrarily close to Perfect Safety. In particular, we will show that in environments that satisfy a standard single-crossing condition, Safe Nash Implementation will be possible for any (Maskin) Monotonic SCF in the *Almost Perfectly Safe* sense (i.e., for all $\epsilon > 0$, (A, k)-Safe Nash Implementation is possible for an acceptability correspondence that satisfies the condition in point 4 of Ex. 2). Also, we stress that the negative result above holds for SCF, but Perfectly Safe Nash Implementation may be achieved if the SCC is non-single valued. The following example illustrates the point:

Example 5 Consider an environment with two states, three alternatives, and four agents, denoted respectively as $\Theta = \{L, R\}$, $X = \{a, b, c\}$, $N = \{1, 2, 3, 4\}$. Preferences are as follows: In state L, players 1 and 2 prefer a to b to c, while players 3 and 4 prefer b to c to a. In state R players 1 and 2 prefer c to b to a, while players 3 and 4 prefer a to c to b. The designer wishes to implement a SCC that selects the alternatives that are at the top of at least half of the agents (hence, $F(L) = \{a, b\}$ and $F(R) = \{a, c\}$), but ensuring *perfect safety*, in the sense that only the outcomes consistent with the SCC are deemed acceptable (that is, $A(L) = \{a, b\} = F(L)$ and $A(R) = \{a, c\} = F(R)$.) Fig. 3 summarizes as usual agents' preferences, the SCC, and the acceptability correspondence. As it will follow from Theorem 3 in the next section, such a SCC can be *perfectly safe* implemented. To see this, first notice that the intersection of player 3's lower contour set of b at state L with the acceptable allocations at that state, are not a subset of his lower contour set at state R. Hence, comonotonicity does not require that $b \in F(R)$. Similarly, comonotonicity does not require that $c \in F(L)$, even if $c \in F(R)$, because the relevant contour set of player 1 at state L is not a subset of that at state R. Indeed, it will be easy to verify that this environment satisfies the sufficient conditions that we provide within the next section, and hence the result will follow directly from Theorem 3. □

Theorem 1 follows directly from the next result, which describes a structural property that must hold for any mechanism that safely implements the SCC. To this end, for any mechanism $\mathcal{M}$, for any $k \geq 1$, and for any $\theta \in \Theta$, let $R_k(\theta) = \bigcup_{m^* \in \mathcal{C}^{\mathcal{M}}(\theta)} B_k(m^*)$, where $\mathcal{C}^{\mathcal{M}}(\theta)$ denotes the set of Nash equilibria of $G^{\mathcal{M}}(\theta)$. That is, $R_k(\theta)$ consists of all message profiles that, given $\mathcal{M}$, are within k deviations from some Nash equilibrium at state θ. Finally, given an acceptability correspondence $A^* : \Theta \to 2^X \setminus \emptyset$ and $k \geq 1$, we say that a mechanism $\mathcal{M} = ((M_i)_{i \in N}, g)$ is *k-surjective on A^** if, for every $\theta \in \Theta$, $g(R_k(\theta)) = A^*(\theta)$. With this, we can finally state the following result:

Theorem 2 (On the Structure of Safe Mechanisms) *Any mechanism that (A, k)-Safe Nash Implements F must be k-surjective on some weakly Comonotonic sub-correspondence of A. If, moreover, A is* maximally *safe, then the implementing mechanism is k-surjective on A itself.*

[14] With unrestricted mechanisms and complete information, the ease with which undesirable equilibria can be ruled out is the main driver of necessity results, more than ensuring non-emptiness of the solution concept.

[15] Shoukry (2019) obtains a slightly weaker version of Corollary 1, in that $A(\theta) = \{f(\theta)\}$ is required at all states as opposed to some.

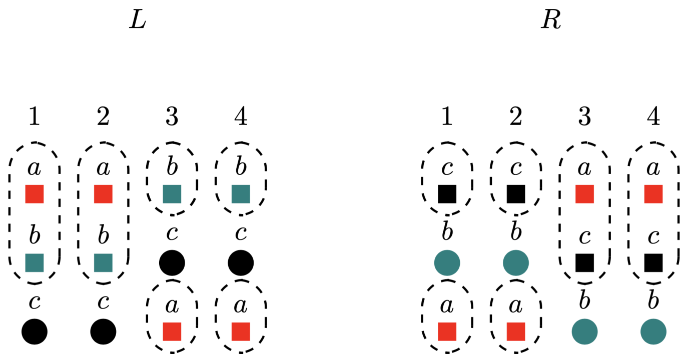

Figure 3: Players 1, 2, 3 and 4's preference orderings over the three alternatives, at the two states, L and R. For each state, the allocation chosen by SCC F in Ex. 5 is indicated by a square. The acceptability correspondence A is shown by the dashed lines, and is *perfectly safe*, as it coincides with the SCC at every state.

Theorem 2 provides a structural property of any implementing mechanism that ties together the restrictions on the acceptability correspondence imposed by weak Comonotonicity, with the *safety level* parameter k. First, this result says that if a mechanism (A, k)-Safely Nash Implements F, then the A^k correspondence defined as $A^k(\theta) := g(R_k(\theta))$ for all $\theta \in \Theta$ is *weakly Comonotonic* and a sub-correspondence of A. This directly implies that A^k and F are weakly Comonotonic, and hence Theorem 1 follows from Theorem 2 when $A^k = A$, as well as the following further necessary condition for (non-maximal) Safe Implementation:

Corollary 2 *A SCC, $F : \Theta \to 2^X \setminus \emptyset$, is (non-maximally) (A, k)-Safe Nash Implementable* only if *A admits a sub-correspondence, A^* such that (A^*, F) satisfy part 2 of Def. 7.*[16]

We note, however, that a non-maximally safe acceptability correspondence may still satisfy part 2 of Def. 7, i.e. with A^* in Corollary 2 be such that $A^*(\theta) = A(\theta)$ for all θ. Consider the following example:

Example 6 To see that we may have a non-maximally safe acceptability correspondence that still satisfies all properties of comonotonicity, consider the following example. Let everything be the same as the leading example, except at state L player 3's preference ordering is $c \succ a \succ b$. Let the SCF be $f^*(L) = f^*(H) = a$, $f^*(M) = b$, with $A(L) = A(M) = \{a, b\}$, and $A(H) = \{a, b, c\}$, as in the leading example. First note that, while it can be shown that f^* can be Safely implemented with respect to A, this acceptability correspondence is not *maximally safe*, since f^* can also be safely implemented with respect to the subcorrespondence A^*, such that $A^*(\theta) = \{a, b\}$ for all θ. Figure 4 summarizes as usual agents' preferences, the SCC, and the two acceptability correspondences. Nonetheless, we show that (A, f^*), in this case, satisfies both conditions for comonotonicity. Part 1 of Def. 7 can be checked following the same logic as in the earlier examples (and it also follows from Proposition 1). To see that part 2 of Def. 7 also holds, note that it cannot be violated due to moving from states L or M to any other state, as $A(L) = A(M) \subseteq A(H)$, and therefore the condition is satisfied regardless. To see there is no violation moving from state H to state L, notice

[16] Putting Proposition 1 and Corollary 2 together, the following is also true: $F : \Theta \to 2^X \setminus \emptyset$, is (non-maximally) (A, k)-Safe Nash Implementable only if A admits a weakly Comonotonic subcorrespondence.

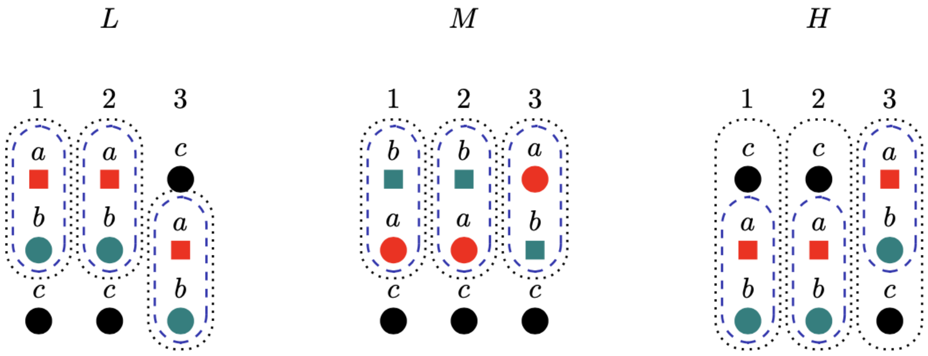

Figure 4: 1, 2 and 3's preference orderings over the three alternatives, at the three states, L, M, and H. For each state, the allocation chosen by SCF f^* in Ex. 6 is indicated by a square. The acceptability correspondence A from this example is shown by the dotted lines, and satisfies the conditions of Weak Comonotonicity. Acceptability correspondence A^* such that $A^*(\theta) = \{a, b\} \subseteq A(\theta)$ is maximally safe, and is represented by the dashed lines in the figure.

that relative to $f^*(H) = a$, an acceptable allocation at H, c, moves up in the ranking of player 3 from state H to L. Therefore we conclude that $A(H) \not\subseteq A(L)$ does not violate part 2 of Def. 7. To see that there is no violation moving from state H to state M, notice that relative to $f^*(H) = a$, an acceptable allocation at H, namely b, has moved up in player 1's ranking from state H to state M. With this, we also conclude that we do not violate part 2 of Def. 7 by setting $A(H) \not\subseteq A(L)$. Hence, A is not maximally safe, and yet it is comonotonic with respect to f^*. □

Finally, notice that holding a mechanism $\mathcal{M}$ fixed, increasing k (weakly) enlarges the set $g(R_k(\theta))$ of outcomes that are within k deviations from the Nash Equilibria at state θ. As long as the corresponding A^k defined as above is weakly Comonotonic and such that $A^k(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$, then the necessary condition for (A, k)-Safe Nash Implementation is satisfied. But if, as k increases, the A^k correspondence is not a sub-correspondence of A, or not weakly Comonotonic, then $\mathcal{M}$ cannot (A, k)-Safe Nash implement the SCC. In that case, Safe Implementation by $\mathcal{M}$ requires either relaxing the admissibility requirement by making A more inclusive (if A^k is not a sub-correspondence of A, or if it violates part 2 of Def. 7), or to 'reduce' the dependence of the SCC on the state of the world (if A^k violates part 1 of Def. 7). In this sense, the structural properties of any implementing 'safe' mechanism in the statement of Theorem 2 reflect a trade-off between the *safety level* parameter $k \geq 1$, the strictness of the *acceptability correspondence*, and the *responsiveness* of the SCC to the state of the world.

5 Sufficiency

Our sufficiency results rely on the following stronger version of Comonotonicity:

Definition 8 (Strong Comonotonicity) *A SCC, $F : \Theta \to 2^X \setminus \emptyset$, and an acceptability correspondence, $A : \Theta \to 2^X \setminus \emptyset$, are* strongly comonotonic *if they satisfy the following conditions:*

1. [A-Constrained Monotonicity of F] *If $\theta, \theta' \in \Theta$ and $x \in F(\theta)$ are such that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ for all $i \in N$, then $x \in F(\theta')$.*

2. [strongly F-Constrained Monotonicity of A] *If $\theta, \theta' \in \Theta$ are such that $\exists x \in F(\theta)$ s.t. $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ for all $i \in N$, then $A(\theta) \subseteq A(\theta')$.*

First, notice that the difference between *Strong* and *Weak Comonotonicity* (Def. 7) is only in the quantifier of the $x \in X$ in part 2 of the definition: in the weak version, the property $A(\theta) \subseteq A(\theta')$ is only required for states $\theta, \theta' \in \Theta$ in which $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ holds for all $i \in N$ and *for all* $x \in F(\theta)$. In contrast, in Def. 8, this property is required to hold for all $\theta, \theta' \in \Theta$ in which $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ holds for all $i \in N$ and *for some* $x \in F(\theta)$. The latter definition therefore is clearly more demanding in general, except when the SCC is single-valued (that is, when the designer wishes to implement a SCF, $f : \Theta \to X$), in which case the two notions of Comonotonicity coincide.

Our main sufficiency result will show that, together with a generalization of Maskin's No-Veto condition, *Strong Comonotonicity* is sufficient for (A,k) Nash Implementation for general SCC. In the case of SCFs, and under such a generalization of the No-Veto condition, *Comonotonicity* (either Def. 7 or 8) is both necessary and sufficient. We introduce next the notion of Safe No Veto:

Definition 9 (Safe No-Veto) *(F,A) are said to satisfy Safe No-Veto if $x \in F(\theta)$ and $A(\theta) = X$ whenever $x \in X$ and $\theta \in \Theta$ are such that $\exists i \in N, \theta' \in \Theta : \forall j \in N \backslash \{i\}, x \in \operatorname{argmax}_{y \in A(\theta')} u_j(y,\theta)$.*

In words, this property restricts both the SCC and the acceptability correspondence at states θ in which all agents but one agree that a particular allocation $x \in X$ is "best" among the set of allocations $A(\theta')$ that are acceptable at some other state θ'. At any such state, the condition requires that the SCC include such x and that all allocations be acceptable.

First note that, if the safety requirement is vacuous (i.e., if $A(\theta) = X$ for all $\theta \in \Theta$), then Def. 9 coincides with Maskin's no veto condition. In all other cases, the condition is stronger than Maskin's No-Veto for two reasons: first, because it suffices that x be at the top for 'almost everyone' only *within the set* $A(\theta') \subset X$, for some $\theta' \in \Theta$, which is implied by being at the top among *all* allocations in X, as requested by the condition for Maskin's No-Veto; second, because it entails a restriction also on the acceptability correspondence, which is required to be vacuous at least such states θ.

Obviously, Def. 9 has no bite if preferences rule out 'almost unanimity' on any subset of allocations, as is the case in many economic settings, such as the single-crossing environments that we will consider in Section 6, or whenever the acceptability correspondence satisfies the following (weaker) 'economic' condition:

Definition 10 (Economic Restrictions) *The acceptability restrictions are* Economic *if, for all $\theta, \theta' \in \Theta$ and $x \in X$, $\left|\{i \in N : x \in argmax_{y \in A(\theta')} u_i(y,\theta)\}\right| < n-1$.*

Furthermore, as we will explain below, Safe No-Veto is almost necessary. Nonetheless, it ensures a fairly strong sufficiency result, which generalizes Maskin's to account for the Safety concerns:

Theorem 3 (Sufficiency) *If $n \geq 3$, and (F,A) are strongly Comonotonic and satisfy Safe No-Veto, then F is (A,k)-Safe Nash Implementable for all $k \in \mathbb{N} : 1 \leq k < \frac{n}{2}$.*

Since Safe-No Veto holds vacuously under Def. 10, Theorem 3 implies the following:

Corollary 3 *If the acceptability restrictions are 'economic', Strong Comonotonicity of (F, A) is sufficient for F to be (A, k)-Safe Nash Implementable for all $k \in \mathbb{N} : 1 \leq k < \frac{n}{2}$.*

We also recall that, in the special case of SCFs, Def. 7 and 8 coincide, and hence Theorems 1 and 3 directly imply the following:

Corollary 4 (Safe Implementation of SCFs under Economic Restrictions) *Let $f : \Theta \to X$ be such that (f, A) satisfy the Safe No-Veto condition (Def. 9).*[17] *Then: (i) f is maximally (A, k)-Safe Nash implementable only if (f, A) are* Comonotonic*; (ii) (f, A) are* Comonotonic *only if f is (A, k)-Safe Nash implementable for all $k \in \mathbb{N} : 1 \leq k < \frac{n}{2}$.*

In the next subsections we further discuss the Safe No-Veto condition, the sense in which it is almost necessary, and various ways in which it can be weakened or dispensed with.

5.1 Safe-No Veto: Almost Necessity in Unrestricted Domains

As we mentioned, the aspect of Safe No-Veto that selects the allocation is almost necessary in a similar sense to Maskin's No Veto being almost necessary for Nash Implementation. In the case of Maskin, a *unanimity property* is necessary, which requires that if *all* agents agree on an allocation being amongst their most preferred at a given state of the world θ, and is implemented at some state θ', then such an allocation must be implemented at θ. No Veto is very similar to this necessary condition, as it differs from it in requiring that if *all but one* agree on an allocation being amongst their most preferred at θ, then such an allocation must be implemented at θ.

Similarly, a necessary condition analogous to unanimity holds for (A, k)-Safe Nash Implementation, and it involves properties of the implementing mechanism that tie back together with the safety parameter k. Specifically, let $\mathcal{M}$ be a mechanism that (A, k)-Safe Nash Implements F, and take any equilibrium m^* at any state $\theta \in \Theta$. Now examine the outcomes that are consistent with $k - 1$ deviations from that equilibrium. By definition of Safe Implementation, all such outcomes would be within the set $A(\theta)$. Now suppose that (i) at some state θ', *all* agents agree that $x \in X$ is most preferred within $A(\theta)$, and (ii) such x is within $k - 1$ deviations from the m^* equilibrium at θ. Then, it must be that x is selected at θ'. Formally:

Lemma 1 *Fix a mechanism that (A, k)-Safe Nash Implements F, and let m^* be a Nash Equilibrium at θ (hence, it is such that $g(m^*) \in F(\theta)$). If for some $\theta' \in \Theta$ we have $x \in g(B_{k-1}(m^*)) \cap \operatorname{argmax}_{y \in A(\theta)} u_i(y, \theta')$ $\forall i \in N$, then $x \in F(\theta')$.*

This necessary condition differs from Safe No-Veto only in two ways: First, it requires x to be within $k - 1$ deviations from an equilibrium at state θ, which need not be the case for all allocations in $A(\theta)$; Second, it requires unanimity of agents' ranking of x at the top of the set $A(\theta)$, as opposed to all but one agreeing on this top element of $A(\theta)$. In this sense, Safe No-Veto is *almost necessary*, as it almost coincides with the necessary condition above.

The other restrictive aspect of Safe No Veto is the fact that, at states in which such a 'unanimity condition' is satisfied, it requires the acceptability restriction to be vacuous. This concession is necessary when preferences are unrestricted, but it can be significantly weakened or dispensed with under minor restrictions on the preferences domain. We discuss some of these weakenings next.

[17] Which is the case, for instance, if the acceptability restrictions satisfy the *economic condition* in Def.10.

5.2 Weakenings and Dispensability of Safe-No Veto

The conditions for Safe No-Veto are not too restrictive under most standard environments, as it is unusual to have preferences where almost all agents agree. An example of this are the single-crossing environments that we discuss in Section 6, or the result we stated in Corollary 3, which shows that Safe No-Veto holds vacuously and hence it can be dropped for the sufficiency result if the acceptability restrictions satisfy the *economic condition* in Definition 10. Furthermore, if the environment satisfies a very slight weakening of this condition, which requires the number of agents that agree on a best alternative within some set of acceptability allocations to be strictly less than n, rather than $n-1$, then the requirement that $A(\theta) = X$ in Safe No-Veto can be weakened to the much more permissive condition that $A(\theta) \subseteq A(\theta')$. Formally:

Definition 11 (weak Safe No-Veto) (F, A) *are said to satisfy* weak *Safe No-Veto if* $x \in F(\theta)$ *whenever* $x \in X$ *and* $\theta \in \Theta$ *are such that* $\exists i \in N, \theta' \in \Theta : \forall j \in N\backslash\{i\}, x \in \operatorname{argmax}_{y \in A(\theta')} u_j(y, \theta)$.

Definition 12 (No unanimity in A) *An environment satisfies* no unanimity in A *if for all* $\theta, \theta' \in \Theta$ *and* $x \in X$, $\left|\{i \in N : x \in argmax_{y \in A(\theta')} u_i(y, \theta)\}\right| < n$.

Proposition 2 *For any* $n \geq 3$, *if* (F, A) *are strongly Comonotonic, satisfy* no unanimity in A *and* weak *Safe No-Veto, then* F *is* (A, k)*-Safe Nash Implementable for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2} - 1$.

5.2.1 Stochastic Mechanisms

Under mild conditions on the environment, Safe No-Veto can be dropped from the sufficient conditions via the use of a stochastic mechanism. Hence, if stochastic mechanisms are allowed, Strong Comonotonicity is sufficient on its own, which in turn provides a full characterisation for Social Choice Functions.[18] Formally: first assume that $u_i(\cdot, \theta)$ represent von Neumann Morgenstern preferences, and say that a SCC is (A, k)-Safely Nash Implementable by a stochastic mechanism if there exists a (possibly stochastic) mechanism $\mathcal{M} = ((M_i)_{i \in I}, g)$ s.t. $g : M \to \Delta(X)$, that (i) Nash Implements it and (ii) such that, for all θ, for all Nash equilibria m^* of $G^{\mathcal{M}}(\theta)$, and for all $m \in B_k(m^*)$, *supp*$(g(m)) \in A(\theta)$. Then, Strong Comonotonicity is sufficient under the following mild domain restriction:

Definition 13 (No Total Indifference across F and A) (F, A) *satisfy No Total Indifference across* F *and* A *if, for all* $\theta, \theta' \in \Theta$, $x \in F(\theta')$ *and* $y \in A(\theta')\backslash\{x\}$, $\exists i \in N$ *s.t.* $u_i(x, \theta) \neq u_i(y, \theta)$.

Proposition 3 (Safe Implementation via Stochastic Mechanisms: Sufficiency) *Under the condition in Def. 13, for all* $n \geq 3$ *and finite* X, *if* (F, A) *are strongly Comonotonic, then* F *is* (A, k)*-Safe Nash Implementable by a stochastic mechanism for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2} - 1$.

For SCFs, this result immediately implies that comonotonicity (weak or strong) is both necessary and sufficient for Safe Nash Implementation via stochastic mechanisms:

Corollary 5 (Safe Implementation of SCFs via Stochastic Mechanisms) *Let* $n \geq 3$ *and* X *be finite. Under the condition in Def. 13:* f *is maximally* (A, k)*-Safe Nash implementable by a stochastic mechanism only if* (f, A) *are* Comonotonic*; (ii)* (f, A) *are* Comonotonic *only if* f *is* (A, k)*-Safe Nash implementable by a stochastic mechanism for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2} - 1$.

[18] This result is analogous to those in Bochet (2007) and Benoît and Ok (2008), who showed that Maskin Monotonicity is both necessary and sufficient for (non-safe) Nash Implementation if stochastic mechanisms are allowed.

5.2.2 Weak Preferences for 'Correctness'

A beaten path within the implementation literature is to consider behavioral preferences in which agents have weak preferences for 'truthfully' reporting the state and allocation (for similar ideas, see Matsushima (2008), Dutta and Sen (2012), Kartik et al. (2014), and Lombardi and Yoshihara (2020). In this case, we show that even if such preferences are 'weak' in the sense of being lexicographically subordinated to the actual outcome of the mechanism, then Safe Nash implementation can be obtained under a mild Unanimity restriction. Formally:

Definition 14 (Weak Preferences for Correctness) *Consider a mechanism $\mathcal{M}$ with message space $M_i = X \times \Theta \times \mathbb{N}$ for all $i \in N$. Agents have a* weak preferences for correctness *if, for all $i \in N$, $u_i : X \times \Theta \times M_i \to \mathbb{R}$ are such that $u_i(x, \theta, (x, \theta, n)) > u_i(x, \theta, (y, \theta, n)) = u_i(x, \theta, (x, \theta', n)) > u_i(x, \theta, (y, \theta', n))$ when $\theta' \neq \theta$ and $y \neq x$.*

Definition 15 (Unanimity within all Acceptable Allocations) *(F, A) satisfy* Unanimity within all Acceptable Allocations (UAA) *if $y \in F(\theta)$ whenever there exists some $\theta' \in \Theta$ such that, for some $m \in M$, $y \in \text{argmax}_{x \in \bigcup_{\theta' \in \Theta} A(\theta')} u_i(x, \theta, m_i)$ for all $i \in N$.*

Proposition 4 (Sufficiency under Weak Preferences for Correctness) *Under weak preferences for correctness, and for all $n \geq 3$, if (F, A) satisfy UAA, then F is (A, k)-Safe Nash Implementable for all $k \in \mathbb{N} : 1 \leq k < \frac{n}{2} - 1$.*

Note that here Safe Nash implementation obtains under the UAA restriction, even without Comonotonicity. In fact, under weak preferences for correctness even *weak* Comonotonicity is not necessary for Safe Nash Implementation. Hence, despite the weakness of these behavioral preferences (recall that they are lexicographically subordinated to the 'standard' preferences over outcomes), they have a profound impact on the possibility of implementation.

6 Special Environments and Applications

We now turn to two canonical applications of Nash Implementation, and include safety concerns. In the first application we explore implementation of SCFs in environments that satisfy a standard single-crossing condition. In this setting, first we show that Comonotonicity is guaranteed whenever the acceptability correspondence includes at every state an ϵ-neighbourhood of the allocation prescribed by the SCF. Second, we show that this condition is sufficient for Safe Nash Implementation for all $k < \frac{n}{2}$. This means that, in these settings, essentially any SCF can be implemented in the *Almost Perfectly Safe* sense that we discussed in p. 10. We then go on to explore the problem of allocating one unit of an indivisible good. We show that, when there is an appropriate *null allocation* that is included in the acceptability correspondence at all states of the world, Safe Nash Implementation of the efficient SCF is possible. Finally, we also provide some negative results, for both Nash implementation and for general solution concepts, in environments that satisfy a strong but standard 'richness condition' on preferences.

6.1 Environments with Private Goods and Single-Crossing Preferences

For each $i \in \{1, ..., n\}$, let $X_i := \mathbb{R}^2_+$ denote the consumption space, with generic consumption bundle denoted as $x_i = (x_i^1, x_i^2)$, with x_i^g denoting the quantity of good g consumed by agent

i. The space of feasible allocations is denoted by $X \subseteq \times_{i \in N} X_i$, assumed compact and convex, with generic element $x = (x_i)_{i \in N}$, which is sometimes convenient to write as $x = (x_i, x_{-i})$, to separate i's own consumption bundle from the profile of consumption bundles of the others. For each agent i, there is a set of types $\Theta_i = \{\theta_i^1, ..., \theta_i^{l_i}\} \subset \mathbb{R}_+$ that pin down i's preferences over X, labelled so that $\theta_i^1 < ... < \theta_i^{l_i}$. The agents' preferences profiles therefore are pinned down by states $\theta \in \Theta = \times_{i \in N} \Theta_i$. The assumption of *private goods* is reflected in that each agent i's utility over X is constant in x_{-i}, and hence utility functions can be written as $u_i(x_i, \theta_i)$, assumed to be continuously differentiable and strictly increasing in both x_i^1 and x_i^2 for each $\theta_i \in \Theta_i$. Finally, we assume that preferences satisfy a single-crossing condition that requires that agents' marginal rate of substitution between good 1 and good 2 is increasing in θ_i for each i.[19]

Letting $f : \Theta \to X$ denote the SCF, it seems sensible to include in the acceptability correspondence allocations that are sufficiently close to $f(\theta)$ at every $\theta \in \Theta$. (This would be natural, for instance, if the social planner chooses $f(\theta)$ to be in the argmax of its welfare criterion, and if the latter is continuous). Formally, for some $\epsilon > 0$ and neighbourhood $\mathcal{N}_\epsilon(f(\theta)) = \{(x_1, x_2) \in X : d(f(\theta), (x_1, x_2)) < \epsilon\}$, where $d(\cdot, \cdot)$ is the Euclidean distance, we assume that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$. This condition is obviously satisfied if $A(\theta) = \mathcal{N}_\epsilon(f(\theta))$, which would make for an especially demanding acceptability criterion, as ϵ gets smaller.

Lemma 2 *Under the maintained single-crossing condition, if the acceptability correspondence $A : \Theta \to 2^X \setminus \emptyset$ is such that, for some $\epsilon > 0$, we have that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$ for all $\theta \in \Theta$, then for* any SCF s.t. $f(\theta) \in int(X)$ for all $\theta \in \Theta$ then (f, A) satisfies (weak and strong) Comonotonicity*.*

In addition to implying Comonotonicity, we show next that in these environments, this minimal condition on the acceptability correspondence also suffices for Safe Nash Implementation, with no need to invoke any additional restrictions.

Proposition 5 *Suppose that $n \geq 3$, and that the single crossing condition above is satisfied. If (f, A) is such that $f(\theta) \in int(X)$ for all $\theta \in \Theta$ and $\exists \epsilon > 0$ such that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$ for all $\theta \in \Theta$, Θ, then f can be (A, k)-Safe Nash Implemented for any $k < \frac{n}{2}$.*

6.2 Efficient Allocation of an Indivisible Good

A social planner wants to allocate an indivisible good to one of the agents in N, or to no agent. The set of feasible outcomes therefore is $X = N \cup \{\emptyset\}$. Like Eliaz (2002), we assume that the set of states and agents' preferences are such that: (P.1) agents always prefer getting the object themselves than having it assigned to someone else; (P.2) conditional on not obtaining the object, agents always prefer it being assigned to agents with a higher utility, and prefer it not being assigned at all over being assigned to someone other than the highest utility agent; and (P.3) at any state of the world, there is always a single agent with the highest valuation.[20] Finally, we assume that the SCF and the acceptability correspondence are such that: (A.1) the SCF is efficient; (A.2) not assigning

[19] Formally, for any $x_i = (x_i^1, x_i^2)$ and $\theta_i, \theta_i' \in \Theta^i$ such that $\theta_i < \theta_i' : \frac{\partial u_i}{\partial x_i^2}(x, \theta_i) / \frac{\partial u_i}{\partial x_i^1}(x, \theta_i) < \frac{\partial u_i}{\partial x_i^2}(x, \theta_i') / \frac{\partial u_i}{\partial x_i^1}(x, \theta_i')$.

[20] Formally, for all i and θ: (P-1) $u_i(i, \theta) > u_i(j, \theta)$ for all $j \in N \backslash \{i\}$; (P.2) $\forall j, k \in N \backslash \{i\}$, $u_i(j, \theta) > u_i(k, \theta)$ if $u_j(j, \theta) > u_k(k, \theta)$, and $u_i(\emptyset, \theta) > u_i(j, \theta)$ if $j \notin \arg\max_{i \in N} u_i(i, \theta)$; and (P.3) $|\arg\max_{i \in N} u_i(i, \theta)| = 1$.

the object is always acceptable; and (A.3) whenever agent i is the designated winner, some other allocation is also acceptable.[21] Under these assumptions, the following possibility result obtains:

Proposition 6 *If $n \geq 3$ and preferences satisfy assumptions P.1-3, any (f, A) that satisfies assumptions A.1-3 is (A, k)-Safe Nash Implementable for all $k < \frac{n}{2}$.*

As already mentioned, the assumptions on the preferences (P.1-3) are the same as in Eliaz 2002), and they are quite weak. Given the minimality of assumptions A.1-3, this result provides a fairly strong possibility result for Safe Nash Implementation of the efficient SCF in allocative problems of a single indivisible good.

6.3 Environments with 'Rich' Preferences

In this subsection we focus on environments that satisfy the following richness condition, analogous to the *Universal Domain* assumption in Social Choice Theory:

Definition 16 *We say that Θ is 'rich' if for every possible profile $\succ = (\succ_i)_{i \in N}$ of strict preference orderings over X, there exists a state of θ such that $u_i(\cdot, \theta)$ represents $\succ_i$ for all $i \in N$.*

Under this condition, we provide two negative results for Safe Implementation. For the first result, we go back to the general definition of Safe Implementation, for general solution concepts $\mathcal{C}$ (cf. Definition 2), and we consider the *minimal safety guarantee* that we introduced in point 1 of Ex. 2. Under these restrictions, the social planner wishes to ensure that, in the case of deviations from the profiles admitted by the solution concept, no agent receives their least preferred outcome. This is a plausible, seemingly minimal criterion for safety restrictions. Yet, under richness, we obtain the following negative result for general solution concepts:

Proposition 7 *Suppose that Θ is rich, $1 < |X| \leq n$. No SCF is (A, k)-Safe $\mathcal{C}$-Implementable for some $k \geq 1$, if A satisfies the minimal safeguarding guarantee.*

The proof of this result is in the appendix. Its main significance is that, in contrast with what could perhaps be surmised from the previous subsections, Safe Implementation is not a vacuous restriction, regardless of the underlying solution concept. For Safe Nash implementation, this message is further reinforced by the following result, which shows that under the richness condition above, if the SCF is onto (i.e., if for any feasible allocation $x \in X$, there is a state $\theta \in \Theta$ such that $f(\theta) = x$), then the Safety requirement can only hold vacuously:

Proposition 8 *Suppose that Θ is rich, and that the SCF, f, is surjective. Then, f is (A, k)-Safe Nash-Implementable for some $k \geq 1$ only if $A(\theta) = X$ for all θ.*

In words, this result says that onto SCF functions cannot be Safe Nash Implemented in 'rich' preferences environments, unless the acceptability requirements are vacuous (in which case the notion coincides with baseline Nash implementation). The proof is as follows: If it is not the case that $A(\theta) = X$ for some θ, then it must be that some $x \in X$ is not in $A(\theta)$. By surjectivity, there is some state where $x = f(\theta')$, and clearly $x \neq z = f(\theta)$. By richness, there is a state θ'' where

[21] Formally: (A.1) $f(\theta) \in \arg\max_{i \in N} u_i(i, \theta)$ for all $\theta \in \Theta$; (A.2) $\forall \theta \in \Theta$, $\{\emptyset, f(\theta)\} \subset A(\theta)$; and (A.3) For any i, whenever $f(\theta) = i$, $\exists x \neq i, \emptyset$ s.t. $x \in A(\theta)$.

x is the top ranked alternative for all players, while z is second ranked for all players. Hence, by Comonotonicity, it should be that both z and x are chosen by the SCF at θ''. But since $x \neq z$, and we have a SCF, this is a contradiction.

7 Related Literature

The closest paper to ours is Eliaz (2002), who studies an implementation problem imposing the requirement that the mechanism's outcome is not affected by deviations of up to k agents. In that sense, the robustness desideratum in Eliaz is more demanding than ours, as it coincides with the special case of 'perfect safety' (which will be discussed below), in which the acceptability correspondence coincides with the SCC. Another important difference is in the solution concept: in Eliaz (2002)'s k-Fault Tolerant (FT) equilibrium, agents reports are required to be optimal not only at the equilibrium profile, but also at all profiles in which up to k agents have deviated. Thus, the solution concept in Eliaz (2002) is stronger than Nash equilibrium, and more so as k increases, with the implementation notion approaching dominant-strategy implementation as k approaches the number of opponents. This has important implications for the comparison with our approach: first, it may be that a SCF is implementable in the sense of Eliaz (2002) but not Nash Implementable – hence, unlike our notion, k-FT Implementation is not necessarily more demanding than baseline Nash Implementation; second, it may be that fault-tolerant implementation is possible for some k, but not for some smaller k' – hence, unlike our notion, the implementation notion in Eliaz (2002) does not necessarily become more demanding as k increases.

Eliaz (2002)'s seminal contribution also inspired Shoukry (2019), which however maintains Nash equilibrium as a solution concept, but like Eliaz (2002) focuses on the special case of 'perfect safety', in which the implementing mechanism is supposed to induce outcomes consistent with the SCF also in the event that up to k agents deviate.[22] As noted, this implies that the SCF is constant. To allow for more positive results, the author allows for transfers and non-standard preferences. In contrast, in this paper, we follow the standard approach of full implementation, with standard preferences and study SCC that select subsets of the whole space of outcomes.[23]As for the safety requirement, our framework allows a wide range of acceptability correspondences, beyond the case of 'perfect safety', and we insist that *all* equilibria be safe.

Another related paper is Hayashi and Lombardi (2019) on "constrained implementation", which studies Nash implementation in a two-sector economy. Within this setting, there is a mechanism for each sector, each determining the allocation of goods within that sector. But while agents' preferences may display complementarities between the goods, and the social planner's objective is to affect the allocation of both goods, he only has freedom to design the mechanism for one sector, taking the other mechanism as given. The possibility of preference interdependence between the two goods leads to a constraint on the planner's ability to elicit preferences using only the freedom that he has to design the mechanism in one sector. This constraint is akin to

[22] Shoukry (2019) also considers SCCs. In that case, he imposes an even stronger restriction than Eliaz's (2002) and our 'perfect safety', in that he demands that the outcome does not change if up to k agents deviate, not just that it stays within the SCC at that state. The two approaches are thus substantially different. In a working paper (Shoukry, 2014) a distinct special case of our acceptability correspondence is considered, where a number of agents cannot obtain alternatives that are too low in their rankings, which yields an impossibility under rich preferences. To regain positive results, the implementation requirement is then weakened so as to effectively allow some equilibria to not be safe. With this, also this approach is profoundly different from ours, and not nested.

[23] That is, we do not leave dimensions of the outcome space, such as transfers, outside of the SCC's codomain.

our acceptability correspondence because only certain allocations within the fixed sector can be achieved by deviations from a candidate equilibrium. Hayashi and Lombardi (2017) also consider a problem similar to Hayashi and Lombardi (2019), but do so in partial equilibrium, where agents only consider deviations within each sector of the economy, not deviations within multiple sectors.

Postlewaite and Wettstein (1989) and Hong (1995) study continuous implementation in a Walrasian economy. They show that the implementing mechanism can be designed so that the outcome function is continuous, and hence such that small deviations from the equilibrium messages lead to small changes in the allocation. This ensures that, even if all agents misreport, if their messages remain sufficiently close to the equilibrium reports, then the outcome will be *close* in the allocation space, which can also be seen as a special instance of our general *acceptability correspondence.*This, however, does not apply to all implementation problems, as many allocation spaces are not naturally endowed with a non-trivial metric. Furthermore, our notion of Safe Implementation does not require agents' possible deviations to remain close to the equilibrium, but we do require that only a certain number of deviations can occur, while at the same time ensuring safety. More broadly, also the literature on feasible implementation (Postlewaite and Wettstein, 1989; Hong, 1995, 1998) is related to our approach. Specifically, as the allocations that occur upon deviations must be feasible at a given state, and the feasibility constraints in this literature may themselves be state-dependent, the notion of implementation indirectly restricts the allocations that can be used upon deviations, much like our notion of Safe Implementation.[24]

Another strand of literature includes concerns for robustness primarily focusing on changes to the solution concept. For instance, Renou and Schlag (2011) study an implementation problem where agents are unsure about the rationality of others, using a solution concept based on ϵ-minmax regret. Similarly, Tumennasan (2013) studies implementation under quantile response equilibrium, letting the logit parameter approach the perfect rationality benchmark. Barlo and Dalkıran (2021) explicitly model the possibility of preference misspecification, letting the states the SCC is based on not pin down agents' preferences, and pursuing a notion of implementation where agents act a la Nash *for all* preferences that are consistent with the designer's information about the states.[25] In our paper, in contrast, we maintain Nash equilibrium and capture the possibility of mistakes (or preference misspecification) as an extra desideratum, on top of the standard notion of implementation. Bochet and Tumennasan (2022b) also maintain Nash Equilibrium, but add the extra requirement that, in a direct mechanism, not only all non-truthful profiles admit a profitable deviation (as required by baseline Nash implementation), but that deviating to truthful revelation is profitable in such instances. This notion is motivated by *resilience* considerations, and is shown to be equivalent to *secure* implementation of Saijo et al. (2007), where implementation is required to occur with respect to both Nash equilibrium and in Dominant Strategies. A related notion can also be found in De Clippel (2014), in which the designer takes into account that agents may have specific kinds of deviations in mind, related to various behavioral considerations. For further recent approaches to behavioral implementation, see De Clippel et al. (2019), Crawford (2021),

[24] This prevents, for instance, that non-equilibrium messages require the designer to import resources. Hurwicz (1979) and Schmeidler (1980), for example, provide positive results for Nash Implementation, and refinements of Nash, in a Walrasian Economy, but deviations from equilibrium may result in non-feasible allocations.

[25] In that sense, Barlo and Dalkıran (2021) can be seen as an original take on the broader idea of robust implementation, where the types that are relevant for the allocation rule pin down agents' preferences, but not their beliefs, which however matter since implementation is required to be achieved for all beliefs consistent with the designer's information (cf. in Bergemann and Morris (2005, 2009a,b), Ollár and Penta 2017, 2022, 2023).

Kneeland (2022), Barlo and Dalkıran (2022), and Bochet and Tumennasan (2022a).

Finally, while based on an unrelated motivation, our results are also connected with the literature on implementation with evidence (e.g., Kartik and Tercieux (2012); Ben-Porath et al. (2019)), which also enriches the baseline Nash implementation framework with an extra desideratum (in that case, the ability to produce evidence about the state of the world). Similar to our *Comonotonicity*, the main condition in that literature is also a suitably adjusted version of monotonicity.

8 Conclusions

We put forward *Safe Implementation*, a notion of implementation that adds to the standard requirements the restriction that deviations from the baseline solution concept induce outcomes that are *acceptable*. This is modelled by introducing, next to the Social Choice Correspondence (which represents the 'first best' objectives when agents behave in accordance with the solution concept), an Acceptability Correspondence that assigns to each state of the world a set of allocations that are considered acceptable, if a number of agents deviate from the solution concept. This framework generalizes standard notions of implementation (which obtain for the special case in which all allocations are 'acceptable') and can accommodate a variety of questions, including robustness concerns with respect to mistakes in play, model misspecification, behavioral considerations, state-dependent feasibility restrictions, limited commitment, etc.

Robustness concerns for mistakes in play and other behavioral considerations have been considered in the literature, mainly through changes to the solution concept (e.g., Eliaz (2002); Renou and Schlag (2011); Tumennasan (2013); De Clippel (2014), De Clippel et al. (2019), Crawford (2021), etc.) Our approach differs mainly in that we impose restrictions also on the outcomes of players' deviations, and may thus be adopted to capture concerns for misspecification of agents' behavior of any kind, as something which can be superimposed on *any* solution concept, be it 'classical' or 'behavioral'.[26] Besides being able to extend robustness concerns to behavioral concepts, modeling them not as part of a specific solution concept has the further advantage of addressing the frequent critique of behavioral models, of being *ad hoc*: in our approach, the deviations that are the object of *Safety considerations* are unrestricted in their nature, and hence model-free.

Decoupling these concerns from the outcomes of the solution concept, however, raises some challenges: on the one hand, like in the standard approach, the outcomes that ensue from deviations must provide the agents with the incentives to induce socially desirable outcomes, consistent with the criteria that are embedded in the underlying solution concept; on the other hand, our concerns for safety limit precisely the designer's ability to specify such outcomes, and the fact that the acceptable allocations are themselves state-dependent, like the SCC, means that not only must agents be given the incentives to induce socially desirable allocations, but also to reveal which outcomes can be used as punishments to achieve this objective. Our main results, which refer to Nash equilibrium as the underlying solution concept, precisely formalize this interplay: the necessary and sufficient conditions that we provide entail joint restrictions on the structure of the SCC and of the acceptability correspondence, and formally generalize the standard conditions for baseline Nash Implementation (Maskin, 1999). While we also offer some results for general

[26] This way, the model can also be used to accommodate general robustness concerns, to account for the possibility that even a behavioral model, which may have been developed in order overcome certain limitations of "classical" notions, may of course also be misspecified.

solution concepts, that identify substantive limits to the possibility of achieving non-trivial Safety desiderata, a systematic exploration of solution concepts other than Nash equilibrium is beyond the scope of this paper, and provides an interesting direction for future research in this area.

Our framework is also general in the specification of the acceptability correspondence, which can be used to accommodate different special cases, which include: (i) the case of "perfectly Safe implementation", which deems acceptable only the outcomes of the SCC (e.g. Eliaz (2002)); (ii) the case of " almost perfectly Safe implementation", when only outcomes that are arbitrarily close to those in the SCC are acceptable, which provides a connection with the literature on continuous implementation (e.g., Postlewaite and Wettstein (1989); Hong (1995)); (iii) the case in which the acceptability correspondence reflects feasibility constraints, which provides a new link to the classical literature on feasible implementation (e.g., Postlewaite and Wettstein (1989); Hong (1995, 1998)); (iv) minimal guarantees based on a variety of welfare criteria (cf. Ex. 2); (v) the possibility to accommodate issues of limited commitment, when the designer can only commit to carrying through, depending on the state, certain punishments but not others (cf., Ex. 1). But these are only some of the possibilities that can be cast within our framework, and further exploring these or other special cases of the acceptability correspondence, explicitly tailored to address specific concerns in more applied settings, may provide another promising direction for future research.

Finally, as it is customary when conceptual innovations are introduced within the implementation literature, and in order to better focus on the essential features of our approach, we have maintained the complete information assumption and imposed no further restrictions on the implementing mechanisms, other than the safety requirements. Combining safety considerations with incomplete information, or with other restrictions on the class of mechanisms (e.g., Jackson (1992), Ollár and Penta (2017, 2022, 2023), etc.), is yet another direction for future research.

Appendix

A Proofs

Proof of Theorem 1: Suppose that F is (A,k)-Safe Nash Implementable. Further, suppose that it is maximally so. Therefore there is some mechanism $\mathcal{M}$ that (A,k)-Safe Implements F and is such that $A(\theta)=g(\{m\in M|d(m,m^*)\leq k, \quad m^*\in\mathcal{C}^{\mathcal{M}}(\theta)\})$.

We will show that F and A are weakly comonotonic in two steps.

Firstly, we will show that if for some $\theta,\theta'\in\Theta$, if there exists $x\in F(\theta)$ such that $L_i(x,\theta)\cap A(\theta)\subseteq L_i(x,\theta')\cap A(\theta)$ for all $i\in N$, then $x\in F(\theta')$. To do so, take m^* to be a Nash Equilibrium at θ that induces x. Hence $g(m^*)=x\in F(\theta)$. Let $\theta'\in\Theta$ be a state such that $x\notin F(\theta')$. Therefore m^* is not a Nash Equilibrium at θ' and hence $\exists i\in N$, $m_i'\in M_i$ such that $u_i(g(m_i',m_{-i}^*),\theta')>u_i(x,\theta')$. It follows that $g(m_i',m_{-i}^*)\in X\backslash L_i(x,\theta')$ and $g(m_i',m_{-i}^*)\in g(\{m\in M|d(m,m^*)\leq k, \quad m^*\in\mathcal{C}^{\mathcal{M}}(\theta)\})=A(\theta)$. However, as m^* is a NE at θ we have that $g(m_i',m_{-i}^*)\in L_i(x,\theta)\cap A(\theta)$. Therefore it cannot be the case that $L_i(x,\theta)\cap A(\theta)\subseteq L_i(x,\theta')\cap A(\theta)$, a contradiction.

Now we show that if for some $\theta, \theta' \in \Theta$, all $x \in F(\theta)$ are such that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ for all $i \in N$, then $A(\theta) \subseteq A(\theta')$. Suppose that θ and θ' are states such that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta')\ \forall i \in N$ for all $x \in F(\theta)$. Suppose by contradiction that $A(\theta) \not\subseteq A(\theta')$.

Take m^* to be a Nash Equilibrium at θ that induces $x \in F(\theta)$. Divide the problem into two cases.

1. m^* is a Nash Equilibrium at θ': in this case we conclude that $B_k(m^*) \subseteq A(\theta')$ by definition.
2. m^* is not a Nash Equilibrium at θ'. In this case, there must be some $i \in N$, who at the state θ' has a profitable deviation from m^*, i.e. $u_i(g(m'_i, m^*_{-i}), \theta') > u_i(x, \theta')$. We conclude that $g(m'_i, m^*_{-i}) \in X \backslash L_i(x,\theta')$. By (A,k)-Safe Nash Implementation, and by definition we have that $A(\theta) = g(\{m \in M | d(m,m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\})$, it must be that $g(m'_i, m^*_{-i}) \in L_i(x,\theta) \cap A(\theta)$. A contradiction to $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta')$ for all $x \in F(\theta)$.

We conclude that all m^* that induce x are Nash Equilibria at θ are also Nash Equilibria at θ'. Now notice that if this holds for all $y \in F(\theta)$ then all Nash Equilibria at θ are also Nash Equilibria at θ'. Given this, the outcomes induced by k agents misreporting from Equilibrium at θ are also reached within k deviations of an Equilibrium at θ'. Concluding that $A(\theta) \subseteq A(\theta')$.

Therefore, we conclude that (F, A) are weakly comonotonic. ■

Proof of Proposition 1:

Suppose that F is (A,k)-Safe Nash Implementable. Therefore there is some mechanism $\mathcal{M}$ that (A,k)-Safe Implements F . We will show that if for some $\theta, \theta' \in \Theta$, if there exists $x \in F(\theta)$ such that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ for all $i \in N$, then $x \in F(\theta')$. That is, A-Constrained Monotonicity of F is satisfied. To do so, take m^* to be a Nash Equilibrium at θ that induces x. Hence $g(m^*) = x \in F(\theta)$. Let $\theta' \in \Theta$ be a state such that $x \notin F(\theta')$. Therefore m^* is not a Nash Equilibrium at θ' and hence $\exists i \in N$, $m'_i \in M_i$ such that $u_i(g(m'_i, m^*_{-i}), \theta') > u_i(x, \theta')$. It follows that $g(m'_i, m^*_{-i}) \in X \backslash L_i(x,\theta')$ and $g(m'_i, m^*_{-i}) \in g(\{m \in M | d(m,m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\}) \subseteq A(\theta)$ by definition of Safety. However, as m^* is a NE at θ we have that $g(m'_i, m^*_{-i}) \in L_i(x,\theta) \cap A(\theta)$. Therefore it cannot be the case that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$. ■

Proof of Theorem 2:

Suppose that F is (A,k)-Safe Nash Implementable. Therefore there is some mechanism $\mathcal{M}$ that (A,k)-Safe Implements F and is such that $g(\{m \in M | d(m,m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\}) \subseteq A(\theta)$. Take A^* to be a sub-correspondence of A such that $g(\{m \in M | d(m,m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\}) = A^*(\theta)$ for all states. By definition, $\mathcal{M}$ is k-surjective on A^*. Moreover, for maximal safety, we would require that $A^*(\theta) = A(\theta)$ for all θ, else some alternatives could be removed, contradicting maximally safe.

With this, the logic of theorem 1 holds exactly, as the proof only relies on the outcomes obtainable within k deviations of the implementing mechanism. That is, one could replace $A(\theta)$ with $A^*(\theta)$ throughout. ■

Proof of Theorem 3: Let each agent $i \in N$ announce an outcome that is acceptable at some state, a state, and a natural number. Thus $M_i = \bigcup_{\theta' \in \Theta} A(\theta') \times \Theta \times \mathbb{N}$, with a typical element $m_i = (x^i, \theta^i, n^i)$. Let $g(m)$ be as follows:

Rule (i) If $m_i = (x, \theta, n^i)$ $\forall i \in N$ and $x \in F(\theta)$ then $g(m) = x$

Rule (ii) If $m_i = (x, \theta, n^i)$ $\forall i \in N \backslash \{j\}$ with $x \in F(\theta)$ and $m_j = (y, \cdot, \cdot)$ then

$$g(m) = \begin{cases} y & \text{if } y \in L_j(x, \theta) \cap A(\theta) \\ x & \text{if } y \notin L_j(x, \theta) \cap A(\theta) \end{cases}$$

Rule (iii) if $k > 1$ and $m_i = (x, \theta, \cdot)$, $x \in F(\theta)$, $\forall i \in N \backslash D$, $2 \leq |D| \leq k$ such that $\forall j \in D$ $m_j \neq (x, \theta, \cdot)$

$$g(m) = \begin{cases} x^{i^*} & \text{if } D^*(\theta, D) \neq \emptyset \\ x & \text{if } D^*(\theta, D) = \emptyset \end{cases}$$

where

$$D^*(\theta, D) = \{j \in D | x^j \in A(\theta)\}$$

and $i^* = \min\{i \in D^*(\theta, D) | n^i \geq n^j \quad j \in D^*(\theta, D)\}$

Rule (iv) Otherwise, let $g(m) = x^{i^*}$ where $i^* = \min\{i \in N | n^i \geq n^j \quad \forall j \in N\}$

From here we can complete the proof in three steps: showing that all $x \in F(\theta)$ are induced by a Nash Equilibrium at θ, showing that there is no $y \notin F(\theta)$ such that y is induced by an Equilibrium at θ, and finally showing that the mechanism is indeed (A, k)-Safe.

Step 1. First to show that all $x \in F(\theta)$ are induced by Nash Equilibria at θ. Consider m^* such that $m_i^* = (x, \theta, \cdot)$, $\forall i \in N$ where $x \in F(\theta)$ at the state θ. To be a Nash Equilibrium we need to rule out the possibility that $\exists j \in N, m_j' \in M_j$ such that $u_j(g(m_{-j}^*, m_j'), \theta) > u_j(g(m^*), \theta)$.

However, $g(m_{-j}^*, m_j') = y$ must be such that $y \in L_j(x, \theta)$ by rule (ii), therefore it is not possible that $u_j(y, \theta) > u_j(x, \theta)$. Given this, m^* is a Nash Equilibrium leading to $x \in F(\theta)$.

Step 2. We will now show that no m^* a Nash equilibrium at θ that is a such that $g(m^*) = y \notin F(\theta)$. We proceed by showing that in each section of the rule, no Nash equilibrium leads to $y \notin F(\theta)$.

Case 1. Suppose m^* is a Nash equilibrium in rule i) at state θ such that $g(m^*) = y \notin F(\theta)$. It must be that $m_i^* = (y, \theta', n^i)$ for all $i \in N$ and, necessarily as $y \notin F(\theta)$, that $\theta' \neq \theta$. Given this, it must be that there is no profitable deviation as m^* is a Nash equilibrium. As deviations may only lead to rule (ii), it must be that for all $i \in N$, for any $z \in L_i(y, \theta') \cap A(\theta')$ we have that $z \in L_i(y, \theta)$, as there is no profitable deviation to report $m_i = (z, \theta, \cdot)$ inducing outcome z from rule (ii). With this, $L_i(y, \theta') \cap A(\theta') \subseteq L_i(y, \theta) \cap A(\theta')$. Therefore, by strong comonotonicity, we have that $y \in F(\theta)$, a contradiction.

Case 2. Now suppose that there is a Nash equilibrium m^*, which is in rule (ii), at state θ such that $g(m^*) = y \notin F(\theta)$. It must be that $\exists j \in N$ such that, $\forall i \in N \backslash \{j\}$ we have $m_i^* = (x, \theta', n^i)$, while $m_j^* \neq (x, \theta', \cdot)$. For this to be a Nash equilibrium it must be that there is not an incentive for any agent to deviate. If $k > 1$ a deviation can lead to rule (i), (ii), or (iii), regardless, as m^* is a Nash equilibrium at θ, no agent $i \neq j$ to wish to change their report, inducing rule (iii), it must be that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$. By Safe No-Veto, it must therefore be that $y \in F(\theta)$, a contradiction to $y \notin F(\theta)$. For $k = 1$ we have that a deviation can lead to rule (i), (ii), or (iv), which in the case of rule (iv) can induce any outcome. Those that can deviate to impose rule (iv) are all agents other than j. With this, we have that, as there is no incentive to deviate, that $y \in \operatorname{argmax}_{z \in \bigcup_{\theta'' \in \Theta} A(\theta'')} u_i(z, \theta)$ for all $i \in N \backslash \{j\}$. With this, it must be that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N \backslash \{j\}$, and therefore by Safe No-Veto we have that $y \in F(\theta)$, a contradiction.

Case 3. Now suppose that there is a Nash equilibrium m^*, which is in rule (iii), at state θ and $g(m^*) = y \notin F(\theta)$. Suppose that $|D| < k$ and $m_i^* = (x, \theta', \cdot)$ for all agents $i \notin D$. Given this, it must be that there is no profitable deviation for any agent. As there exists a message for any player that leads to any allocation in $A(\theta')$ via rule (iii), we conclude that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N$. Therefore by Safe No-Veto, we have that $y \in F(\theta)$. Now suppose that $|D| = k$. For there to be no profitable deviation, it must be that for $\forall i \in D$, $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$. For all agents in $i \in N \backslash D$ it must be that for any $x \in \bigcup_{\theta'' \in \Theta} A(\theta'') \supseteq A(\theta\prime)$, we have that $u_i(y, \theta) \geq u_i(x, \theta)$, as there is no profitable deviation. Given this, we conclude that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N$, and therefore by Safe No-Veto we conclude that $y \in F(\theta)$, a contradiction.

Case 4. Finally, if there is a Nash equilibrium m^* at θ in rule (iv), we can see that a unilateral deviation can lead to any outcome in $\bigcup_{\theta'' \in \Theta} A(\theta'')$ via rule (iv). With this, it must be that for m^* with $g(m^*) = y$ to be a Nash equilibrium in this state we have that $y \in \operatorname{argmax}_{z \in \bigcup_{\theta'' \in \Theta} A(\theta'')} u_i(z, \theta)$ for all $i \in N$. Therefore, $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for some θ', and therefore by Safe No-Veto we have that $y \in F(\theta)$.

Step 3. We will now show that all Nash equilibria are safe. To do so, we will again split it into cases.

Case 1. If m^* is a Nash equilibrium at θ that falls into rule (i) it must be that $m_i^* = (y, \theta', n^i)$. By the previous analysis, we know that $y \in F(\theta)$. If $\theta' = \theta$, we conclude that safety is satisfied as k deviations can only lead to rule (ii) or (iii). Either way, we remain in $A(\theta)$. Now suppose that $\theta' \neq \theta$ while m^* is a Nash equilibrium at θ. Notice that regardless, k deviations must lead to remaining within $A(\theta')$ via rule (ii) or (iii). By the previous analysis, we know that this only occurs when $L_i(y, \theta') \cap A(\theta') \subseteq L_i(y, \theta) \cap A(\theta')$ for all $i \in N$. Given this, $A(\theta') \subseteq A(\theta)$ must hold for strong comonotonicity to be satisfied. Given that $A(\theta') \subseteq A(\theta)$, we conclude that any deviation from such a Nash equilibrium must remain in $A(\theta')$, and therefore $A(\theta)$, maintaining safety.

Case 2. Now suppose that m^* is a Nash equilibrium at θ that falls into rule (ii). It must be that $\forall i \neq j$ $m_i^* = (x, \theta', n^i)$ while $m_j^* \neq (x, \theta', n^i)$. Notice that k deviations can lead to rule (i), rule (iii) if $k > 1$, and rule (iv). Notice that in fact, k deviations can lead to rule (iii) for some state $\theta'' \neq \theta'$ if $k = \frac{n}{2} - 1$, depending on the report of j. Regardless, safety will require that $A(\theta) = \bigcup_{\theta'' \in \Theta} A(\theta'')$ for this mechanism. To see this is implied by the condition of Safe No-Veto, notice that by the previous analysis, we only have a Nash equilibrium at such a state if $\forall i \notin N \backslash \{j\}$ we have that they prefer to stick to $g(m^*) = y$ rather than inducing any outcome in rule (iii), in the case $k > 1$, or rule (iv), in the case, that $k = 1$. Given this, it must be that $y \in \text{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \notin N \backslash \{j\}$. Given this, it must be that by Safe No-Veto we have that $A(\theta) = \bigcup_{\theta'' \in \Theta} A(\theta'')$, concluding that safety is maintained.

Case 3. Now suppose that m^* is a Nash equilibrium at θ that falls into rule (iii), and therefore $k > 1$. It must be that all agents in $i \in N \backslash D$ for some $D \subset N$ with $|D| \leq k$, are reporting $m_i^* = (x, \theta', n^i)$. Notice that by the structure of the mechanism, k deviations can lead to:

1. rule (i), leading $x \in F(\theta')$ only by those in D changing their report.
2. rule (ii) it is possible to reach rule (ii) if $|D| - 1$ people report $m_j = (x, \theta', \cdot)$.
3. if $n = 3$, while $k = 1$, it could also be that we lead to rule (ii) at another state due to a change in the report of one agent who isn't the whistle-blower.
4. rule (iii) leading to any outcome in $A(\theta')$ by those in D or, if $|D| < k$, this may also be due to $k - |D|$ agents within $N \backslash D$ changing their report.
5. If all those in D report $m_j = (z, \theta'', n^j)$, as well as $k \geq |D| > \frac{n}{4}$, it could also be that k deviations lead to rule (ii) causing acceptable allocations at state θ'' to be reached by k agents in $N \backslash D$ changing their report to (z, θ'', n^i).
6. rule (iv) by $k - |D| + 1$ of those who are in $N \backslash D$ changing reports.

With this, it is possible that for safety to be achieved we require that $A(\theta) = \bigcup_{\theta''} A(\theta'')$. Notice that for $y = g(m^*)$ to be a Nash equilibrium at state θ, by the previous analysis it must be that $y \in \text{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N$. With this, it must then be that by safety no veto $A(\theta) = \bigcup_{\theta'' \in \Theta} A(\theta'')$. Therefore Safety is necessarily achieved.

Case 4. Finally, suppose that m^* is a Nash equilibrium at θ with $g(m^*) = y$. Note that by the rules of the mechanism, k deviations can lead to

1. Any outcome via rule (iv)
2. if $m^* = (x, \theta', n^i)$ for sufficiently many agents then it could also lead to any other rule for state θ'.

If we have a Nash equilibrium within this rule, it must be that $y \in \text{argmax}_{z \in X} u_i(z, \theta)$ for all $i \in N$, as else any agent could deviate to induce any outcome in $\bigcup_{\theta'' \in \Theta} A(\theta'')$ they wish via announcing a higher integer. With this, we conclude that it must be that $y \in \text{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for any

$A(\theta')$ such that $y \in A(\theta')$. With this, by Safe No Veto, we conclude that $A(\theta) = \bigcup_{\theta'' \in \Theta} A(\theta'')$, and therefore Safety is achieved. ■

Proof of lemma 1:

Suppose for some $m^{\star,\theta}$ we have that $x \in g(B_{k-1}(m^{\star,\theta})) \cap \operatorname{argmax}_{y \in A(\theta)} u_i(y, \theta')\ \forall i \in N$. As $x \in g(B_{k-1}(m^{\star,\theta}))$ it follows that $\exists D_{k-1} \subset N_{k-1}, m_{D_{k-1}} \in M_{D_k}$ with $g(m_{D_{k-1}}, m^{\star,\theta}_{-D_{k-1}}) = x$.

Any unilateral deviation leads to an allocation in $A(\theta)$ by definition of (A,k)-Safe implementation and less than k agents are reporting a non-Equilibrium message. Therefore $m_{D_{k-1}}, m^{\star,\theta}_{-D_{k-1}}$ is a Nash Equilibrium at θ' and therefore $g(m_{D_{k-1}}, m^{\star,\theta}_{-D_{k-1}}) \in F(\theta')$. ■

Proof of Proposition 2: Let each agent $i \in N$ announce an outcome that is acceptable at some state, a state, and a natural number. Thus $M_i = \bigcup_{\theta'' \in \Theta} A(\theta'') \times \Theta \times \mathbb{N}$, with a typical element $m_i = (x^i, \theta^i, n^i)$. Let $g(m)$ be as follows:

Rule (i) If $m_i = (x, \theta, n^i)\ \forall i \in N$ and $x \in F(\theta)$ then $g(m) = x$

Rule (ii) If $m_i = (x, \theta, n^i)\ \forall i \in N \backslash \{j\}$ with $x \in F(\theta)$ and $m_j = (y, \cdot, \cdot)$ then

$$g(m) = \begin{cases} y & \text{if } y \in L_j(x,\theta) \cap A(\theta) \\ x & \text{if } y \notin L_j(x,\theta) \cap A(\theta) \end{cases}$$

Rule (iii) $m_i = (x, \theta, \cdot)$, $x \in F(\theta)$, $\forall i \in N \backslash D$, $2 \leq |D| < \frac{n}{2}$ such that $\forall j \in D\ m_j \neq (x, \theta, \cdot)$

$$g(m) = \begin{cases} x^{i^*} & \text{if } D^*(\theta, D) \neq \emptyset \\ x & \text{if } D^*(\theta, D) = \emptyset \end{cases}$$

where

$$D^*(\theta, D) = \{j \in D | x^j \in A(\theta)\}$$

and $i^* = \min\{i \in D^*(\theta, D) | n^i \geq n^j \quad j \in D^*(\theta, D)\}$

Rule (iv) Otherwise, let $g(m) = x^{i^*}$ where $i^* = \min\{i \in N | n^i \geq n^j \quad \forall j \in N\}$

From here we can complete the proof in three steps: showing that all $x \in F(\theta)$ are induced by a Nash Equilibrium at θ, showing that there is no $y \notin F(\theta)$ such that y is induced by an Equilibrium at θ, and finally showing that the mechanism is indeed (A,k)-Safe.

Step 1. First to show that all $x \in F(\theta)$ are induced by Nash Equilibria at θ. Consider m^* such that $m^*_i = (x, \theta, \cdot), \quad \forall i \in N$ where $x \in F(\theta)$ at the state θ. To be a Nash Equilibrium we need to rule out the possibility that $\exists j \in N, m'_j \in M_j$ such that $u_j(g(m^*_{-j}, m'_j), \theta) > u_j(g(m^*), \theta)$.

However, $g(m^*_{-j}, m'_j) = y$ must be such that $y \in L_j(x, \theta)$ by rule (ii), it is not possible that $u_j(y, \theta) > u_j(x, \theta)$. Therefore it must be that m^* is a Nash Equilibrium leading to $x \in F(\theta)$.

Step 2. We will now show that no m^* a Nash equilibrium at θ that is a such that $g(m^*) = y \notin F(\theta)$. We proceed by showing that in each section of the rule, no Nash equilibrium leads to $y \notin F(\theta)$.

Case 1. Suppose m^* is a Nash equilibrium in rule i) at state θ such that $g(m^*) = y \notin F(\theta)$. It must be that $m_i^* = (y, \theta', n^i)$ for all $i \in N$ and, necessarily as $y \notin F(\theta)$, that $\theta' \neq \theta$. Given this, it must be that there is no profitable deviation and therefore, as deviations may only lead to rule (ii), it must be that for all $i \in N$, for any $z \in L_i(y, \theta') \cap A(\theta')$ we have that $z \in L_i(y, \theta)$, as there is no profitable deviation to report $m_i = (z, \theta, \cdot)$ inducing outcome z from rule (ii). With this, $L_i(y, \theta') \cap A(\theta') \subseteq L_i(y, \theta) \cap A(\theta')$. Therefore, by strong comonotonicity, we have that $y \in F(\theta)$, a contradiction.

Case 2. Now suppose that there is a Nash equilibrium m^*, which is in rule (ii), at state θ such that $g(m^*) = y \notin F(\theta)$. It must be that $\exists j \in N$ such that, $\forall i \in N \backslash \{j\}$ we have $m_i^* = (x, \theta', n^i)$, while $m_j^* \neq (x, \theta', \cdot)$. For this to be a Nash equilibrium it must be that there is no incentive for any agent to deviate. If $k > 1$ a deviation can lead to rule (i), (ii), or (iii), regardless, as m^* is a Nash equilibrium at θ, no agent $i \neq j$ to wish to change their report, inducing rule (iii), it must be that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$. By weak Safe No-Veto, it must therefore be that $y \in F(\theta)$, a contradiction to $y \notin F(\theta)$. For $k = 1$ we have that a deviation can lead to rule (i), (ii), or (iv), which in the case of rule (iv) can induce any outcome. Those that can deviate to impose rule (iv) are all agents other than j. With this, we have that, as there is no incentive to deviate, that $y \in \operatorname{argmax}_{z \in \bigcup_{\theta'' \in \Theta} A(\theta'')} u_i(z, \theta)$ for all $i \in N \backslash \{j\}$. With this, it must be that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N \backslash \{j\}$, and therefore by weak Safe No-Veto we have that $y \in F(\theta)$, a contradiction.

Case 3. Notice that there can be no Nash equilibria within rule (iii). Suppose that m^* were a Nash equilibrium in rule (ii) at state θ. Suppose that $|D| < \lfloor \frac{n}{2} \rfloor$ and $m_i^* = (x, \theta'. \cdot)$ for all agents $i \notin D$. Given this, it must be that there is no profitable deviation for any agent. As there exists a message for any player that leads to any allocation in $A(\theta')$ via rule (iii), we conclude that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N$. Therefore it must be that no unanimity in A is violated. Now suppose that $|D| = \lfloor \frac{n}{2} \rfloor$. For there to be no profitable deviation, it must be that for $\forall i \in D$, $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$. For all agents in $i \in N \backslash D$ it must be that for any $x \in X \supseteq A(\theta')$, we have that $u_i(y, \theta) \geq u_i(x, \theta)$, as there is no profitable deviation. Given this, we conclude that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N$, and therefore no unanimity in A is violated.

Case 4. Note that there can be no Nash equilibria within rule (iv). To see this, suppose that m^* is a Nash equilibrium at state θ that falls within rule (iv), with $g(m^*) = y$. Notice that any agent can deviate to remain within rule (iv), inducing any outcome that is acceptable at any state. Therefore for y to be a Nash equilibrium it must be that $y \in \operatorname{argmax}_{z \in \bigcup_{\theta'' \in \Theta} A(\theta'')} u_i(z, \theta)$ for all $i \in N$. Therefore it follows that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for some $\theta' \in \Theta$ and for all $i \in N$. Therefore no unanimity in A is violated.

Step 3.

We will now show that all Nash equilibria are safe. To do so, we will again split it into cases.

By the previous analysis, recall that if we maintain No Unanimity in A we know that there can only be equilibria in rule (i) or rule (ii), and therefore we need only focus on the safety of those equilibria in rules (i) and (ii).

Case 1. If m^* is a Nash equilibrium at θ that falls into rule (i) it must be that $m_i^* = (y, \theta', n^i)$. By the previous analysis, we know that $y \in F(\theta)$. If $\theta' = \theta$, we conclude that safety is satisfied as k deviations can only lead to rule (ii) or (iii). Either way, we remain in $A(\theta)$. Now suppose that $\theta' \neq \theta$ while m^* is a Nash equilibrium at θ. Notice that regardless, k deviations must lead to remaining within $A(\theta')$ via rule (ii) or (iii). By the previous analysis, we know that this only occurs when $L_i(y, \theta') \cap A(\theta') \subseteq L_i(y, \theta) \cap A(\theta')$ for all $i \in N$. Given this, $A(\theta') \subseteq A(\theta)$ must hold for strong comonotonicity to be satisfied. Given that $A(\theta') \subseteq A(\theta)$, we conclude that any deviation from such a Nash equilibrium must remain in $A(\theta')$, and therefore $A(\theta)$, maintaining safety.

Case 2. Now suppose that m^* is a Nash equilibrium at θ that falls into rule (ii). It must be that $\forall i \neq j$ $m_i^* = (x, \theta', n^i)$ while $m_j^* \neq (x, \theta', n^i)$. Notice that k deviations can lead to rule (i), rule (ii) or rule (iii), as $k < \lfloor \frac{n}{2} \rfloor - 1$. Notice that by the structure of the mechanism, even with k, in the extreme case where $\frac{n}{2} - 2$ misreports from m^*, it remains that the majority of agents are reporting $m_i = (x, \theta', n^i)$. With this, any k deviations must lead to $A(\theta')$. Notice that for this to be a Nash equilibrium at θ, we therefore require that $g(m^*) = y \in \text{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \neq j$. With this, by Weak Safe No-Veto, we have that $A(\theta') \subseteq A(\theta)$. As k deviations remain in $A(\theta')$ it is also true that k deviations remain in $A(\theta)$. Therefore Safety is upheld. ■

Proof of Proposition 3: Let each agent $i \in N$ announce an outcome, a state, and a natural number. Thus $M_i = X \times \Theta \times \mathbb{N}$, with a typical element $m_i = (x^i, \theta^i, n^i)$. Let $g(m)$ be as follows:

Rule (i) If $m_i = (x, \theta, n^i)$ $\forall i \in N$ and $x \in F(\theta)$ then $g(m) = x$

Rule (ii) If $m_i = (x, \theta, n^i)$ $\forall i \in N \backslash \{j\}$ with $x \in F(\theta)$ and $m_j = (y, \cdot, n^j)$ then

$$g(m) = \begin{cases} \frac{n^j}{n^j+1} y + \frac{1}{n^j+1} x & \text{if } y \in L_j(x, \theta) \cap A(\theta) \\ x & ify \notin L_j(x, \theta) \cap A(\theta) \end{cases}$$

Rule (iii) if $m_i = (x, \theta, \cdot)$, $x \in F(\theta)$, $\forall i \in N \backslash D$, $2 \leq |D| \leq \frac{n}{2}$ such that $\forall j \in D$ $m_j \neq (x, \theta, \cdot)$

$$g(m) = \begin{cases} \sum_{y \in A(\theta)} \frac{1}{|A(\theta)| + \sum_{k \in D^*(\theta, D)} n^k} y + \sum_{j \in D^*(\theta, D)} \frac{n^j}{|A(\theta)| + \sum_{k \in D^*(\theta, D)} n^k} x^j & \text{if } D^*(\theta, D) \neq \emptyset \\ \sum_{y \in A(\theta)} \frac{1}{|A(\theta)|} y & \text{if } D^*(\theta, D) = \emptyset \end{cases}$$

where

$$D^*(\theta, D) = \{j \in D | x^j \in A(\theta)\}$$

.

Rule (iv) Otherwise, let $g(m) = \sum_{x \in X} \frac{1}{|X| + \sum_{j \in N} n^j} x + \sum_{i \in N} \frac{1}{|X| + \sum_{j \in N} n^j} x^i$.

From here we can complete the proof in three steps: showing that all $x \in F(\theta)$ are induced by a Nash Equilibrium at θ, showing that there is no $y \notin F(\theta)$ such that y is induced by an Equilibrium at θ, and finally showing that the mechanism is indeed (A, k)-Safe.

Step 1. First to show that all $x \in F(\theta)$ are induced by Nash Equilibria at θ. Consider m^* such that $m_i^* = (x, \theta, \cdot), \quad \forall i \in N$ where $x \in F(\theta)$ at the state θ. To be a Nash Equilibrium we need to rule out the possibility that $\exists j \in N, m_j' \in M_j$ such that $u_j(g(m_{-j}^*, m_j'), \theta) > u_j(g(m^*), \theta)$.

By rule (ii), the only way that $g(m_{-j}^*, m_j') \neq x$, i.e. not to give the deterministic allocation x, it must be that it puts positive weight on x and on one other allocation $y \in L_j(x, \theta) \cap A(\theta)$. Given that $y \in L_j(x, \theta)$, there is no profitable deviation.

Step 2. We will now show that no m^* a Nash equilibrium at θ that is such that $g(m^*) \notin F(\theta)$, i.e. no Nash equilibrium gives anything but the deterministic allocations of $F(\theta)$. We proceed by showing that in each section of the rule, no Nash equilibrium leads to any $y \notin F(\theta)$, or any probabilistic allocation.

Case 1. Suppose m^* is a Nash equilibrium in rule i) at state θ such that $g(m^*) = y \notin F(\theta)$. It must be that $m_i^* = (y, \theta', n^i)$ for all $i \in N$ and, necessarily as $y \notin F(\theta)$, that $\theta' \neq \theta$. Given this, it must be that there is no profitable deviation, and therefore, as deviations may only lead to rule (ii), it must be that for all $i \in N$, for any $z \in L_i(y, \theta') \cap A(\theta')$ we have that $z \in L_i(y, \theta)$, as there is no profitable deviation to report $m_i = (z, \theta, \cdot)$ inducing outcome z from rule (ii). With this, $L_i(y, \theta') \cap A(\theta') \subseteq L_i(y, \theta) \cap A(\theta')$. Therefore, by strong comonotonicity, we have that $y \in F(\theta)$, a contradiction.

Case 2. Now suppose that there is a Nash equilibrium m^*, which is in rule (ii), at state θ such that $g(m^*) \notin F(\theta)$. It must be that $\exists j \in N$ such that, $\forall i \in N \backslash \{j\}$ we have $m_i^* = (x, \theta', n^i)$, while $m_j^* \neq (x, \theta', \cdot)$. We split this possibility into sub-cases for clarity.

Case 2.a. First consider the case that $g(m^*) = x \in F(\theta')$. It must therefore be that for all $i \neq j$ there is no profitable deviation. Given this, we must have that $u_i(x, \theta) \geq \max_{z \in A(\theta') \backslash \{x\}} u_i(z, \theta)$ by the fact a deviation to announce an arbitrarily high n^i, therefore, inducing a probabilistic outcome putting almost probability 1 on their most preferred outcome z. Further, for j to have no profitable deviation we have that it must be that there is no $y \in L_j(x, \theta') \cap A(\theta')$ such that $u_j(y, \theta) > u_j(x, \theta)$. Therefore for all $y \in L_j(x, \theta') \cap A(\theta')$ we have that $u_j(x, \theta) \geq u_j(y, \theta)$, and therefore it is the case that $L_j(x, \theta') \cap A(\theta') \subseteq L_j(x, \theta) \cap A(\theta')$. Further, we have that $L_i(x, \theta) \cap A(\theta') = A(\theta')$ for all $i \neq j$. With this $L_i(x, \theta') \cap A(\theta') \subseteq L_i(x, \theta) \cap A(\theta')$ for all $i \in N$. Therefore by Strong Comonotonicity we have that $x \in F(\theta)$ and $A(\theta') \subseteq A(\theta)$.

Case 2.b. Now instead consider the case where $g(m^*) = \frac{n^j}{n^j+1} y + \frac{1}{n^j+1} x$. As for all $\theta, \theta' \in \Theta$, for all $z \in F(\theta')$ $z' \in A(\theta')$, $\exists i \in N$ such that $u_i(z, \theta) - u_i(z', \theta) \neq 0$, it must be that case that agent is such that $u_i(y, \theta) - u_i(x, \theta) \neq 0$. If such agent is j, i.e. the whistle blower, then a profitable deviation exists to announce either a higher n^j, putting more weight on y, or announce $m_j' = m_i^*$ for $i \neq j$, putting weight 1 on x. Now suppose that $u_i(y, \theta) - u_i(x, \theta) \neq 0$ for $i \neq j$, while $u_j(y, \theta) - u_i(x, \theta) = 0$. Firstly, suppose that $u_i(y, \theta) > u_i(x, \theta)$. Notice that $\forall \epsilon > 0$ $\exists n^i \in \mathbb{N}$

such that $\epsilon > \frac{|A(\theta')|}{|A(\theta')|+n^i+n^j}(u_i(y,\theta) - \min_{z\in A(\theta')} u_i(z,\theta))$. Therefore, simply rearranging this, we have that $\forall\epsilon > 0\ \exists n^i \in \mathbb{N}$ such that $\frac{n^i+n^j}{|A(\theta')|+n^i+n^j}u_i(y,\theta) + \frac{|A(\theta')|}{|A(\theta')|+n^i+n^j}\min_{z\in A(\theta')} u_i(z,\theta) > u_i(y,\theta) - \epsilon$. Given this, we conclude that $\forall\epsilon > 0\ \exists n^i \in \mathbb{N}$ such that $\frac{n^i+n^j}{|A(\theta')|+n^i+n^j}u_i(y,\theta) + \sum_{z\in A(\theta')} \frac{1}{|A(\theta')|+n^i+n^j}u_i(z,\theta) > u_i(y,\theta) - \epsilon$. Let $\epsilon = u_i(y,\theta) - \frac{n^j}{n^j+1}u_i(y,\theta) - \frac{1}{n^j+1}u_i(x,\theta)$. By assumption that $u_i(y,\theta) - u_i(x,\theta) > 0$ we have that $\epsilon > 0$. With this, $\exists n^i \in \mathbb{N}$ such that $\frac{n^i+n^j}{|A(\theta')|+n^i+n^j}u_i(y,\theta) + \sum_{z\in A(\theta')} \frac{1}{|A(\theta')|+n^i+n^j}u_i(z,\theta) > \frac{n^j}{n^j+1}u_i(y,\theta) + \frac{1}{n^j+1}u_i(x) = u_i(g(m^*),\theta)$. With this, announcing $m'_i = (y,\theta,n^i)$ induces such an outcome $u_i(g(m'_i, m^*_{-i}),\theta) = \frac{n^j}{n^j+1}u_i(y,\theta) - \frac{1}{n^j+1}u_i(x) > u_i(g(m^*),\theta)$ and therefore m^* cannot be an equilibrium. By an analogous argument, there cannot be an equilibrium if $u_i(x,\theta) > u_i(y,\theta)$ for some agent, as they can announce an arbitrarily high n^i and x, putting almost probability 1 on x. Regardless, this m^* such that $g(m^*) = \frac{n^j}{n^j+1}y + \frac{1}{n^j+1}x$ cannot be an equilibrium.

Case 3 and 4. Note that there cannot be any equilibria in rule (iii) or rule (iv). To see this, notice that any agent can announce their most one of their most preferred outcome from $A(\theta)$, in rule (iii), or X in rule (iv), and an integer higher than any other agent (including themselves before the deviation), and strictly increase their utility by reducing the probability assigned to their less preferred option. As at least one agent is not completely indifferent between all allocations by No total indifference across F and A, one such agent always exists.

Step 3. Notice that by the previous analysis, there may on be equilibria in rules (i) and (ii), therefore we need only check the Safety of such equilibria.

Case 1. Firstly, suppose that m^* is a Nash equilibrium in rule (i) at state θ. By the previous analysis, we know that it is the case that $m^*_i = (x,\theta',\cdot)$, with $x \in F(\theta')$. If $\theta' \neq \theta$, then, by the previous analysis, we know that $L_i(x,\theta') \cap A(\theta') \subseteq L_i(x,\theta) \cap A(\theta')$ for all $i \in N$. Therefore by Strong Comonotonicity we have that $x \in F(\theta)$ and $A(\theta') \subseteq A(\theta)$. Now notice that in k deviations from m^*, we may only reach $A(\theta')$, via rule (ii), with 1 deviation, or rule (iii) which can be reached with more than 1 but less than $k+1$ deviations. As $k < \frac{n}{2} - 1$, it is the case that the majority of agents still report m^*_i, regardless of what the other k report. Given that $A(\theta') \subseteq A(\theta)$, it follows that any allocation with k deviations of m^* is still a mix with a support of $A(\theta)$. If instead m^* is such that $\theta' = \theta$, Safety is upheld as k deviations can only lead to stochastic allocations over $A(\theta)$. Therefore Safety is upheld.

Case 2. Now instead suppose m^* is a Nash equilibrium at state θ that falls into rule (ii). It must be that $m^*_i = (x,\theta',\cdot)$ for all $i \neq j$ and $m^*_j = (y,\theta'',\cdot)$. By the previous analysis, we know that $g(m^*) = x$. If $\theta' \neq \theta$, again by the previous analysis, we know that it must be that $L_i(x,\theta') \cap A(\theta') \subseteq L_i(x,\theta) \cap A(\theta')$ for all $i \in N$. Therefore by strong comonotonicity we have that $x \in F(\theta)$ and $A(\theta') \subseteq A(\theta)$. Now notice that in k deviations we may reach rule (i) inducing x, rule (ii) inducing mixes over allocations in $L_j(x,\theta') \cap A(\theta')$, or rule (iii) for inducing stochastic allocations over $A(\theta')$. Notice that no other allocations can be reached in k deviations as $k < \frac{n}{2} - 1$, and therefore the majority of agents would still be reporting m^*_i. With this, and by $A(\theta') \subseteq A(\theta)$, we have that the mechanism is still considered Safe. Similarly, if $\theta' = \theta$, we have that in k deviations we may reach rule (i) inducing x, rule (ii) inducing mixes over allocations in $L_j(x,\theta) \cap A(\theta)$, or rule (iii) for inducing stochastic allocations over $A(\theta)$. With this, Safety is upheld.■

Proof of proposition 4:

Take the mechanism and logic to be similar to that of theorem 3:

Let each agent $i \in N$ announce an outcome that is acceptable at some state, a state, and a natural number. Thus $M_i = \bigcup_{\theta'' \in \Theta} A(\theta'') \times \Theta \times \mathbb{N}$, with a typical element $m_i = (x^i, \theta^i, n^i)$. Let $g(m)$ be as follows:

Rule (i) If $m_i = (x, \theta, n^i)$ $\forall i \in N$ and $x \in F(\theta)$ then $g(m) = x$

Rule (ii) If $m_i = (x, \theta, n^i)$ $\forall i \in N \backslash \{j\}$ with $x \in F(\theta)$ and $m_j = (y, \theta', \cdot)$ then

$$g(m) = \begin{cases} y & \text{if } u_i(x, \theta, (x, \theta, \cdot)) \geq u_i(y, \theta, (y, \theta', \cdot)) \text{ and } y \in A(\theta) \\ x & \text{if either } u_i(x, \theta, (x, \theta, \cdot)) < u_i(y, \theta, (y, \theta', \cdot)) \text{ or } y \notin A(\theta) \end{cases}$$

Rule (iii) If $m_i = (x, \theta, \cdot)$, $x \in F(\theta)$, $\forall i \in N \backslash D$, $2 \leq |D| < \frac{n}{2}$ such that $\forall j \in D$ $m_j \neq (x, \theta, \cdot)$

$$g(m) = \begin{cases} x^{i^*} & \text{if } D^*(\theta, D) \neq \emptyset \\ x & \text{if } D^*(\theta, D) = \emptyset \end{cases}$$

where

$$D^*(\theta, D) = \{j \in D | x^j \in A(\theta)\}$$

and $i^* = \min\{i \in D^*(\theta, D) | n^i \geq n^j \quad j \in D^*(\theta, D)\}$

Rule (iv) Otherwise, let $g(m) = x^{i^*}$ where $i^* = \min\{i \in N | n^i \geq n^j \quad \forall j \in N\}$

From here we can complete the proof in three steps: showing that all $x \in F(\theta)$ are induced by a Nash Equilibrium at θ, showing that there is no $y \notin F(\theta)$ such that y is induced by an Equilibrium at θ, and finally showing that the mechanism is indeed (A, k)-Safe. We will proceed by showing that all Nash Equilibria are contained in rule (i), and report the correct state, and therefore, in comparison to theorem 3, we may weaken Safe No-Veto to only Unanimity within all acceptable allocations.

Step 1. First to show that all $x \in F(\theta)$ are induced by Nash Equilibria at θ. Consider m^* such that $m^*_i = (x, \theta, \cdot)$, $\quad \forall i \in N$ where $x \in F(\theta)$ at the state θ. To see this is a NE, suppose not there is some agent for which there is a profitable deviation m'_j, $g(m^*_{-j}, m'_j) = y$ must be such that $u_i(x, \theta, (x, \theta, n)) \geq u_i(y, \theta, (y, \theta', n))$ and $y \in A(\theta)$ (or it is not profitable) by rule (ii), a contradiction to $u_j(y, \theta, (y, \theta', n)) > u_j(x, \theta, (x, \theta, n))$. Therefore it must be that m^* is a Nash Equilibrium leading to $x \in F(\theta)$.

Step 2. We will now show that no m^* a Nash equilibrium at θ that is a such that $g(m^*) = y \notin F(\theta)$. We proceed by showing that in each section of the rule, no Nash equilibrium leads to $y \notin F(\theta)$.

Suppose m^* is a Nash equilibrium in rule i) at state θ such that $g(m^*) = y \notin F(\theta)$. It must be that $m^*_i = (y, \theta', n^i)$ for all $i \in N$ and, necessarily as $y \notin F(\theta)$, that $\theta' \neq \theta$. However, consider a deviation for player i to a report of $m_i = (y, \theta', \cdot)$. This induces the outcome y still. By the definition of weak preference for correctness, we have that $u_i(y, \theta', (y, \theta', \cdot)) > u_i(y, \theta', (y, \theta, \cdot))$, and therefore a profitable deviation exists. A contradiction that m^* being an equilibrium.

Suppose that we have an Equilibrium in case (ii) with $m^*_i = (x, \theta, n^i)$ for all $i \neq j$ and $m^*_j = (y, \cdot, \cdot)$. Suppose the true state is θ'. For this to be the case, it must be that no agent has an incentive to deviate. Therefore it must be that $g(m^*) \neq x$, as otherwise j has an incentive to deviate by announcing $m_j = (x, \theta, n^j)$, and by a preference for correctness would now be announcing the correct state and / or allocation that the mechanism implements. Therefore it must be that $g(m^*) = y$. However, given this, any agent $i \neq j$ has the incentive to deviate to $m_i = (y, \theta', n^i)$, and therefore leading to allocation y via rule (iii) or rule (iv). Via the weak preference for correctness, this strictly increases utility. Therefore there can be no equilibria in rule (ii).

Suppose that the Equilibrium m^* at state θ' is in rule (iii), with $m^*_i = (x, \theta, \cdot)$, $x \in F(\theta)$, $\forall i \in N \backslash D$, $2 \leq |D| < \frac{n}{2}$ such that $\forall j \in D$ $m^*_j \neq (x, \theta, \cdot)$. It must be that at least $|D|$ agents are such that they are either reporting the wrong state or not reporting the allocation that is being implemented $g(m^*) = y$, be that those in D or those in $N \backslash D$. Given this, consider one such agent. They may report the allocation y and/or the state θ' and an integer higher than any other agent. To see this does not change the allocation first consider the case $\frac{n}{2} > |D| > 2$. In such a case, we remain in rule (iii) or rule (iv) via this deviation, where the deviating agent is announcing the highest integer and therefore y is allocated. Now consider the case where $|D| = 2$. First consider $g(m^*) = y = x$. Suppose that $\theta \neq \theta'$, then any agent in $N \backslash D$ may deviate to announce $m_i = (x, \theta', n^i)$ with n^i being higher than any integer announced under m^*. As this announcement announces the true state, it strictly increases the utility of i. Therefore it cannot be that m^* is a Nash equilibrium in this case. Suppose instead that $\theta = \theta'$. Then it must be that those in D are either:

1. Both announcing an allocation not in $A(\theta)$, in which case a deviation by either to $m_j = (x, \theta, \cdot)$ would not change the allocation but would make the report correct, therefore increasing their utility.

2. One is announcing an allocation in $A(\theta)$, while one is not. In which case, there is at least one who is not announcing x, in which case they can increase their utility by doing so.

3. Both are announcing allocations in $A(\theta)$. If this is the case, if both announce x it must be that both are announcing $\theta^j \neq \theta$, and therefore can improve their utility by announcing $m_j = (x, \theta, \cdot)$, and increase their utility, leading to rule (ii), but keeping the same allocation. Now suppose that only one is announcing x. It must be that the other is not, and therefore can increase their utility by announcing x, while keeping the other parts of the report the same, strictly increasing their utility. If neither is announcing x, it cannot be that $g(m^*) = x$.

Now instead consider $g(m^*) = y \neq x$. In such a case, those outside of D may deviate to announce $m_i = (y, \theta', \cdot)$, increasing their utility.

Finally, consider the possibility of an equilibrium m^* in rule (iv) at state θ leading to the outcome y. For this to be the case, it must be that there is no incentive to deviate. Consider the possibility that $m_i^* \neq (y, \theta, \cdot)$ for some i. For this to be the case, it must be that announcing $(y, \theta, \cdot)$ and an integer higher than any other announced under m^* would change the allocation, as otherwise, the preference for correctness would mean a profitable deviation occurs. This can only occur if such a deviation would lead to rule (iii), i.e. $\lfloor \frac{n}{2} \rfloor - 1$ agents are reporting $(x, \theta', \cdot)$. Given this, we can deduce that $(x, \theta', \cdot) = (y, \theta, \cdot)$ and that $y \in F(\theta)$, as otherwise, rule (iv) would dictate the allocation remains the same, while at least one of those $\lfloor \frac{n}{2} \rfloor - 1$ agents could strictly increase their utility by announcing $m_j = (y, \theta, \cdot)$. With this, the original player i such that $m_i^* \neq (y, \theta, \cdot)$ has a profitable deviation as they can announce $(y, \theta, \cdot)$ and some arbitrarily high n^i, inducing rule (iii), where y is chosen due to i announcing the highest integer and $y \in F(\theta) \subseteq A(\theta)$. Therefore it cannot be that $m_i^* \neq (y, \theta, \cdot)$ for any i in any equilibria in rule (iv). For such equilibria to fall into rule (iv) rather than rule (i), it must be that $y \notin F(\theta)$. However, for there to be no profitable deviation within this rule it must therefore be that $y \in \text{argmax}_{x \in \bigcup_{\theta'' \in \Theta} A(\theta'')} u_i(x, \theta, m_i)$ for all $i \in N$, and therefore by Unanimity within all Acceptable Allocations we have that $y \in F(\theta)$. ■

Proof of lemma 2:

Take $\theta, \theta' \in \Theta$ such that $f(\theta) = x \neq f(\theta')$. Let agent i be such that $\theta_i \neq \theta'_i$. Without loss of generality, suppose that $\theta'_i > \theta_i$. We need to show $\exists y \in A(\theta)$ such that $y \in L_i(f(\theta), \theta)$ while $y \notin L_i(f(\theta), \theta')$. By Taylor's theorem, $\exists \epsilon > 0$ such that for $\mathcal{N}_\epsilon(x)$ the remainder term of the 1 Taylor expansion is sufficiently small to preserve inequalities. Therefore we need to show that there exists $y \in \mathcal{N}_\epsilon(x)$ such that $(y_1^i - x_1^i)\frac{\partial u_i(f(\theta),\theta_i)}{\partial x_1^i} + (y_2^i - x_2^i)\frac{\partial u_i(f(\theta),\theta_i)}{\partial x_2^i} < 0$ while $(y_1^i - x_1^i)\frac{\partial u_i(f(\theta),\theta'_i)}{\partial x_1^i} + (y_2^i - x_2^i)\frac{\partial u_i(f(\theta),\theta'_i)}{\partial x_2^i} > 0$ as $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$. With some rearranging we find $\frac{\frac{\partial u_i(f(\theta),\theta^i)}{\partial x_2^i}}{\frac{\partial u_i(f(\theta),\theta_i)}{\partial x_1^i}} < -\frac{y_1^i - x_1^i}{y_2^i - x_2^i} < \frac{\frac{\partial u_i(f(\theta),\theta'_i)}{\partial x_2^i}}{\frac{\partial u_i(f(\theta),\theta'_i)}{\partial x_1^i}}$, which as $\theta'_i > \theta_i$ is satisfied by single crossing, as we can find $-\frac{y_1^i - x_1^i}{y_2^i - x_2^i}$ satisfying the inequalities needed in the neighbourhood. ■

Proof of proposition 5:

Let each agent $i \in N$ announce an outcome, which excludes all reports that would be their maximal allocation, and the state. Therefore $M_i = int(X) \times \Theta$, with typical element $m_i = (x(i), \theta(i))$ Let $g(m)$ be as follows:

Rule (i) If $m_i = (x(i), \theta(i))$ is such that $\theta(i) = \theta \quad \forall i \in N$ then $g(m) = f(\theta)$.

Rule (ii) If $m_i = (x(i), \theta(i))$ is such that $\theta(i) = \theta \quad \forall i \in N \backslash \{j\}$ where $m_j = (x(j), \theta')$, $\theta' \neq \theta$

$$g(m) = \begin{cases} x(j) & \text{if } x(j) \in L_j(f(\theta), \theta) \cap \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta)) \\ f(\theta) & \text{if } x(j) \notin L_j(f(\theta), \theta) \cap \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta)) \end{cases}$$

Rule (iii) If $\exists D \subset N$ such that $k \geq |D| > 1$, where $m_i = (x(i), \theta(i))$ and $\theta(i) = \theta, \forall i \in N \backslash D$, then $g(m)$ is constructed by the following: Let ϵ be fixed across agents such that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$. $\forall i \in D$ let $\tilde{x}(i) = x(i)$ if $x(i) \in \mathcal{N}_{\frac{\epsilon}{|D|}}(f(\theta))$. $\tilde{x}(i) = \lambda^i x(i) + (1 - \lambda^i) f(\theta)$ such that $d(f(\theta), \tilde{x}(i)) = \frac{\epsilon}{|D|+1}$, $\lambda^i \in (0, 1)$ otherwise. where Now let $g(m) = f(\theta) + \sum_{i \in D}(\tilde{x}(i) - f(\theta))$.

Rule (iv) Otherwise, let $g(m) = \frac{1}{n}\sum_{i\in N} x(i)$.

Step 1. First to show that $x = f(\theta)$ is a Nash Equilibrium at θ. Consider m^* satisfying rule (i) Any unilateral deviation of agent i leads to rule (ii), where the only way to change the allocation is in $L_i(f(\theta), \theta)$, which cannot give a strictly higher utility by definition. Therefore all m^* satisfying rule (i) are Equilibria.

Step 2. We want to show that $\nexists m^*$ such that m^* is an Equilibrium at θ such that $g(m^*) \neq f(\theta)$.

Case 1: Suppose that there is an Equilibrium in Rule (i) where $g(m^*) \neq f(\theta)$, where the true state is θ. It follows that all agents are announcing some state $\theta' \neq \theta$. With this, there exists some agent who announces their own type to be $\theta_j(j) = \theta'_j \neq \theta_j$. For this agent $\exists x_j$ s.t. $x_j \in L_j(f(\theta'), \theta') \cap \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta'))$ while $x_j \notin L_j(f(\theta'), \theta) \cap \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta'))$ by the same logic as lemma 2 via the single crossing condition. Therefore m^* cannot be a Nash Equilibrium.

Case 2: we can not have any Nash Equilibria for any state θ in rule (ii). Suppose that m^* is an equilibrium at that θ where for all $i \in N\backslash\{j\}$ we have that $m_i = (x(i), \theta(i))$ with $\theta(i) = \theta'$ while $m_j = (x(j), \theta(j))$ with $\theta(j) \neq \theta'$. Regardless of whether $g(m^*) = f(\theta)$ or $g(m^*) = x(j)$, notice that any agent $i \neq j$ can induce an increase in both dimensions of the bundle by announcing $m_i = (x'(i), \theta'(i))$, where $\theta'(i) \neq \theta'$ and $x'(i)$ such that $x'_j(i) = f_j(\theta)$ and $x'_i(i)$ is chosen such that $x'(i) \in \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta))$ and $\frac{x_i^{k,\prime}(i) + \tilde{x}_j^k(j)}{2} > f_i^k(\theta)$, which is achievable by the construction of rule (iii). As u_i is strictly increasing this is a contradiction to m^* being a Nash equilibrium.

Case 3: There cannot be an Equilibrium in Rule (iii), any agent $i \in D$ can announce an allocation to the north east of $\tilde{x}(i)$ such that $x(i) \in \mathcal{N}_{\frac{\epsilon}{|D|}}(f(\theta))$, leading to rule (iii) or (iv), regardless, monotonically increase their allocation (notice this is the case due to the penalty for announcing $x(i) \notin \mathcal{N}_{\frac{\epsilon}{|D|}}(f(\theta))$ of moving closer to the original Equilibrium $x = f(\theta)$).

Case 4: The final case is within rule (iv). Again, this cannot be an Equilibrium. Agents can continue to announce an allocation to the north east of the current one, leading to rule (iv) except increasing their payoff by the assumption of increasing utility. As the message can only be interior in X, such a message is always feasible as a maximal element cannot be selected.

Step 3: Notice all Equilibria lie in Rule (i). Further, any such equilibrium m^* at θ lead to $g(m^*) = f(\theta)$ by Case 1 of Step 2. With this, deviations of size k or less all lead to to only to rules (i), via announcing a different allocation but same state, rule (ii), via announcing a different state, or rule (iii). k deviations that remains in rule rule (i) must lead to the same allocation, and therefore safety is guaranteed. k deviations that lead to rule (ii) lead to allocations in $\mathcal{N}_{\frac{\epsilon}{2}}(f(\theta)) \subset \mathcal{N}_{\epsilon}(f(\theta)) \subseteq A(\theta)$ and therefore safety is maintained. The only check needed for this is that rule (iii) lies within an ϵ neighbourhood of $f(\theta)$, and therefore within $A(\theta)$. To see this, notice that

$$\begin{aligned}
d(f(\theta), g(m)) &= d\left(f(\theta), f(\theta) + \sum_{i\in D}(\tilde{x}(i) - f(\theta))\right) \\
&= \|\sum_{i\in D}(\tilde{x}(i) - f(\theta))\| \\
&\le \sum_{i\in D}\|\tilde{x}(i) - f(\theta)\| && \text{(by the triangle inequality)} \\
&= \sum_{i\in D} d(f(\theta), \tilde{x}(i)) \\
&< |D|\frac{1}{|D|}\epsilon \\
&= \epsilon
\end{aligned}$$

and therefore $d(f(\theta), g(m)) < \epsilon$, or rather $g(m) \in \mathcal{N}_\epsilon(f(\theta))$ for any m that falls within rule (iii) and is k deviations from an equilibrium at θ. ■

Proof of proposition 6:

Let $X = N \cup \{0\}$, where 0 represents the good being unallocated. For each a state of the world $\theta \in \Theta$ let $\theta \in \mathbb{R}^n_+$, where θ represents vector of values of players.

Let $M_i = X \times \mathbb{R}^n_+$ for all $i \in N$ with a typical message $m_i = (j, v) \in N \cup \{0\} \times \mathbb{R}^n_+$. Let $g(m)$ be as follows:

Rule (i) If $\forall i \in N$ $m_i = (j', v)$ with $v = \theta \in \Theta$ and $j' = f(\theta)$ then $g(m) = j' = f(\theta)$.

Rule (ii) If $m_i = (j', v)$ $\forall i \in N\backslash\{j\}$ with $v = \theta \in \Theta$ and $f(\theta) = i'$ and $m_j = (l, \cdot)$, then

$$g(m) = \begin{cases} l & \text{if } l \in [L_j(j', \theta) \cap \tilde{A}(\theta)]\backslash\{j'\} \\ \emptyset & \text{if } l \notin [L_j(j', \theta) \cap \tilde{A}(\theta)]\backslash\{j'\} \end{cases}$$

Rule (iii) If $m_i = (j', v)$ such that $v = \theta \in \Theta$ and $j' = f(\theta)$ for $\forall i \in N\backslash D$, $2 \le |D| < \frac{n}{2}$ such that $\forall j \in D$ $m_j = (l^j, \cdot)$, $l^j \neq j'$ then

$$g(m) = \begin{cases} l^{i^*} & \text{if } D^*(\theta, D) \neq \emptyset \\ j' & \text{if } D^*(\theta, D) = \emptyset \end{cases}$$

where

$$D^*(\theta, D) = \{j \in D | l^j \in \tilde{A}(\theta)\}$$

and $i^* = \min\{i \in D^*(\theta, D) | v^i_i \ge v^j_j \quad j \in D^*(\theta, D)\}$.

Rule (iv) otherwise let $g(m) = l^{i^*}$ where $m_i = (l^i, \cdot)$ and $i^* = \min\{i \in N | v^i_i \ge v^j_j \quad j \in N\}$.

Where $\tilde{A}(\theta) = \bigcap_{\theta' \in \Theta | f(\theta') = f(\theta)} A(\theta')$.

Notice that, at state θ, with messages that fall into rule (i) with $m^* = (j', \theta)$, m^* is a Nash equilibrium. This is as a deviation either leads to the good being unallocated, which is not preferred to the deserving agent receiving it, or it must be that a less deserving agent receives the good. To show all Nash Equilibria are considered safe, we will do so by showing that rule *i*) constitute the only Nash Equilibria, and always allocate the $f(\theta)$ at state θ.

Suppose that there is a Nash Equilibrium in rule *ii*) m^* at state θ. Let $m_i^* = (j', \theta')$ for all $i \neq j$ and $m_j^* = (l, \cdot)$. It must be either $g(m^*) = l \in \tilde{A}(\theta')$, $l \in N \backslash \{j'\}$, or $g(m^*) = 0$. Suppose that $j = j'$. Here there is a profitable deviation to announce $m_j = (j', \theta')$ and be allocated the good, which cannot be case under rule (ii). Suppose instead that $j \neq j'$. Let $i = j'$, who can announce $m_i = (i, v'')$ such that v_i'' is strictly higher than the i^{th} (or equivalently $j'^{,th}$) component of θ' and receive the good by inducing rule (iii).

As all agents prefer to have the good allocated to themselves, there can be no Equilibria in rule *iii*) and *iv*). To see that in the case of rule (*iii*) there is no Nash equilibrium, suppose that the message of $|N| - k$ agents is $m_i = (j', v')$, with $v' = \theta'$ and $f(\theta') = j'$, while m^* is a Nash equilibrium. Given that there is some agent $j \in \tilde{A}(\theta')$ such that $g(m^*) \neq j$ by (A.3.). Such an agent prefers to have the good allocated to themself, they can announce $m_j = (j, v'')$, such that $v_j'' = \max_{i \neq j} v_i^i + \epsilon$, and therefore would induce that the good is allocated to them. Similarly for rule (*iv*), however now any agent who is not allocated the good could make such a deviation.

Finally, suppose that there is some Nash Equilibrium in rule i) m^* at θ such that, for some θ' we have $g(m^*) = f(\theta') = j' \neq f(\theta)$. It must be that this agent j' is not the agent with the highest valuation, and therefore is undeserving. Given this, any agent can announce $l = 0$ (or any $l \notin A(\theta)$), which given rule (ii) and (P.2.), induces no agent to receive the good, as is not preferred at θ'. However, this is preferred at θ as reverting to the empty allocation is attainable and by assumption gives a higher payoff than an undeserving agent.

Finally, to notice that all Nash equilibria are safe, notice that they all lie within rule (i) with $m_i^* = (j', \theta)$ at state θ', where j' has the highest valuation in state θ'. With this, up to k deviations can only lead to rules (ii) or (iii), where the majority of agents still announces (j', θ). With this, we remain in $\tilde{A}(\theta) \subseteq A(\theta')$, and therefore safety is maintained. ■

Proof of proposition 7:

Proof. If $|X| \leq n$ under the richness condition $\exists \theta \in \Theta$ such that for every $x \in X$ $\exists i \in N$ such that $\{x\} = \text{argmin}_{y \in X} u_i(y, \theta)$. Hence if A is minimally safeguarding then $X^*(\theta) = \emptyset$ and therefore no SCC can be safely $\mathcal{C}$-implemented for any $k \geq 1$ and any $\mathcal{C}$. ■

References

Arya, A., Glover, J., and Rajan, U. (2000). Implementation in principal–agent models of adverse selection. *Journal of Economic Theory*, 93(1):87–109.

Barlo, M. and Dalkıran, N. A. (2021). Implementation with missing data.

Barlo, M. and Dalkıran, N. A. (2022). Behavioral implementation under incomplete information. *Working Paper, Sabanci University.*

Ben-Porath, E., Dekel, E., and Lipman, B. L. (2019). Mechanisms with evidence: Commitment and robustness. *Econometrica*, 87(2):529–566.

Benoît, J.-P. and Ok, E. A. (2008). Nash implementation without no-veto power. *Games and Economic Behavior*, 64(1):51–67.

Bergemann, D. and Morris, S. (2005). Robust mechanism design. *Econometrica*, 73(6):1771–1813.

Bergemann, D. and Morris, S. (2009a). Robust implementation in direct mechanisms. *The Review of Economic Studies*, 76(4):1175–1204.

Bergemann, D. and Morris, S. (2009b). Robust virtual implementation. *Theoretical Economics*, 4(1).

Bergemann, D., Morris, S., and Tercieux, O. (2011). Rationalizable implementation. *Journal of Economic Theory*, 146(3):1253–1274.

Bochet, O. (2007). Nash implementation with lottery mechanisms. *Social Choice and Welfare*, 28(1):111–125.

Bochet, O. and Tumennasan, N. (2022a). One truth and a thousand lies: Defaults and benchmarks in mechanism design. *Working Paper.*

Bochet, O. and Tumennasan, N. (2022b). Resilient mechanisms. *Working Paper.*

Crawford, V. P. (2021). Efficient mechanisms for level-k bilateral trading. *Games and Economic Behavior*, 127:80–101.

De Clippel, G. (2014). Behavioral implementation. *American Economic Review*, 104(10):2975–3002.

De Clippel, G., Saran, R., and Serrano, R. (2019). Level-mechanism design. *The Review of Economic Studies*, 86(3):1207–1227.

Dutta, B. and Sen, A. (2012). Nash implementation with partially honest individuals. *Games and Economic Behavior*, 74(1):154–169.

Eliaz, K. (2002). Fault tolerant implementation. *The Review of Economic Studies*, 69(3):589–610.

Gneezy, U. and Rustichini, A. (2000). A fine is a price. *The journal of legal studies*, 29(1):1–17.

Hayashi, T. and Lombardi, M. (2017). Implementation in partial equilibrium. *Journal of Economic Theory*, 169:13–34.

Hayashi, T. and Lombardi, M. (2019). Constrained implementation. *Journal of Economic Theory*, 183:546–567.

Hong, L. (1995). Nash implementation in production economies. *Economic Theory*, 5(3):401–417.

Hong, L. (1998). Feasible bayesian implementation with state dependent feasible sets. *Journal of Economic Theory*, 80(2):201–221.

Hurwicz, L. (1979). Outcome Functions Yielding Walrasian and Lindahl Allocations at Nash Equilibrium Points. *The Review of Economic Studies*, 46(2):217–225.

Jackson, M. O. (1992). Implementation in undominated strategies: A look at bounded mechanisms. *The Review of Economic Studies*, 59(4):757–775.

Jain, R., Korpella, V., and Lombardi, M. (2022). Two-player rationalizable implementation. *IES Working paper series*, 21(A001).

Jain, R. and Lombardi, M. (2022). On interim rationalizable implementation. *SSRN 4106795*.

Kartik, N. and Tercieux, O. (2012). Implementation with evidence. *Theoretical Economics*, 7(2):323–355.

Kartik, N., Tercieux, O., and Holden, R. (2014). Simple mechanisms and preferences for honesty. *Games and Economic Behavior*, 83:284–290.

Kneeland, T. (2022). Mechanism design with level-k types: Theory and an application to bilateral trade. *Journal of Economic Theory*, 201:105421.

Levitt, S. D. and Dubner, S. J. (2006). Freakonomics: A rogue economist explores the hidden side of everything by.

Lombardi, M. and Yoshihara, N. (2020). Partially-honest nash implementation: a full characterization. *Economic Theory*, 70(3):871–904.

Maskin, E. (1999). Nash equilibrium and welfare optimality. *Review of Economic Studies*, 66(1):23–38.

Maskin, E. and Sjöström, T. (2002). Implementation theory. *Handbook of social Choice and Welfare*, 1:237–288.

Matsushima, H. (2008). Role of honesty in full implementation. *Journal of Economic Theory*, 139(1):353–359.

Mirrlees, J. A. (1976). Optimal tax theory: A synthesis. *Journal of public Economics*, 6(4):327–358.

Moore, J. and Repullo, R. (1988). Subgame perfect implementation. *Econometrica: Journal of the Econometric Society*, pages 1191–1220.

Ollár, M. and Penta, A. (2017). Full implementation and belief restrictions. *American Economic Review*, 107(8):2243–77.

Ollár, M. and Penta, A. (2022). Efficient full implementation via transfers: Uniqueness and sensitivity in symmetric environments. In *AEA Papers and Proceedings*, volume 112, pages 438–43.

Ollár, M. and Penta, A. (2023). A network solution to robust implementation: The case of identical but unknown distributions. *Review of Economic Studies*.

Postlewaite, A. and Wettstein, D. (1989). Feasible and continuous implementation. *The Review of Economic Studies*, 56(4):603–611.

Renou, L. and Schlag, K. (2011). Implementation in minimax regret equilibrium. *Games and Economic Behavior*, 71(2):527–533.

Saijo, T., Sjostrom, T., and Yamato, T. (2007). Secure implementation. *Theoretical Economics*, 2(3):203–229.

Schmeidler, D. (1980). Walrasian analysis via strategic outcome functions. *Econometrica: Journal of the Econometric Society*, pages 1585–1593.

Shoukry, G. (2014). Safety in mechanism design and implementation theory. *Available at SSRN 2478655*.

Shoukry, G. F. (2019). Outcome-robust mechanisms for nash implementation. *Social Choice and Welfare*, 52(3):497–526.

Spence, A. M. (1980). Multi-product quantity-dependent prices and profitability constraints. *The Review of Economic Studies*, 47(5):821–841.

Tumennasan, N. (2013). To err is human: Implementation in quantal response equilibria. *Games and Economic Behavior*, 77(1):138–152.