WORKING PAPERS

N° 1369

September 2022

“Safe Implementation”

Malachy James Gavan and Antonio Penta

Toulouse
School of
Economics

Safe Implementation[*]

Malachy James Gavan[†]
UPF and BSE

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

September 22, 2022

Abstract

We introduce *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*. 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 considerations, 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 (1978), 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, 1978) 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 – each pinning down agents' preferences over the feasible allocations – and a Social Choice Correspondence (SCC) that specifies, for each state, 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

[*] We are grateful to Larbi Alaoui, Alexander Frug and William Sandholm for their comments on early versions of this work.

[†] *email:* malachy.gavan@upf.edu

[‡] *email:* antonio.penta@upf.edu

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 within 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 the 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. 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 is often restricted by other constraints or desiderata, such as constitutional rights, higher level legislation, culture, or social norms.[3]

(ii) An employer wants to promote one of many employees, based on some characteristics (such as motivation, leadership and personal skills) that are known to the employees but not to the employer. However, she wants to ensure that, in the case of a mistake in the employees' reports, it is not the case that she promotes someone with very poor leadership skills, therefore restricting the set of alternatives that may arise from deviations in a way that depends on the state.

(iii) 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, who are highly integrated in the system, but not to the central banker, who however can design a mechanism to elicit it and ensure that the optimal level of stability will be reached in equilibrium. 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.

(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

[1] For instance, Nash Implementation (Maskin, 1978) or Subgame Perfect Implementation (Moore and Repullo, 1988), or more recently Rationalizable (Bergemann et al., 2011), Level-k (De Clippel et al., 2019), and Behavioral (De Clippel, 2014) Implementation. For a thorough survey of the early literature, see Maskin and Sjöström (2002).

[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, 2022a,b), etc., but by and large the existing approaches have not paid special attention to the outcomes that a mechanism may induce, other than at the profiles that are consistent with the solution concept. Two exceptions, albeit with important differences compared to our approach, are 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 and social choice theory have 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 applications, although it did not tackle the aspects that will be the focus of this paper, i.e. the designer's constraints on the outcomes induced by agents' deviations.

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]

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), 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

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

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 *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 it equally with the other firms with which they compete (so, firms 1 and 2 share markets 1 and 2, and each firm gets one third of the market in 3). 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 weak (W), medium (M) or high (H). The true state is known to the firms but not to the designer. Firm 3's ranking is always such that $a \succ b \succ c$. The other firms' preferences over the alternatives instead depend on the state. 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. Their ranking over alternatives at this state therefore is $b \succ a \succ c$. When the demand in country 3 is weak, neither firm 1 or 2 are willing to give up their monopolies in order to access the third market. Hence, their ranking over the alternatives are s.t. $a \succ b \succ c$, the same as firm 3. 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 public 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 $f(W) = a$ 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

[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.

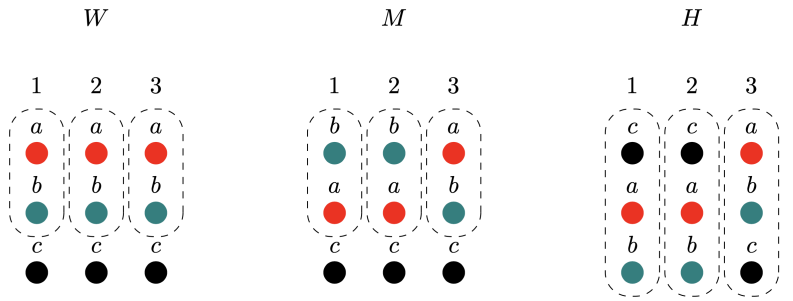

Figure 1: Firms 1, 2 and 3's preference orderings over the three alternatives, at the three states, W, M, and H. The acceptability correspondence, shown in dashed lines, is such that $A(W) = A(M) = \{a, b\}$, $A(H) = \{a, b, c\}$. In this setting, the SCF such that $f(W) = a$ and $f(M) = f(H) = b$ is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

of everyone's preferences, 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 3 go bankrupt, or because it would not be politically credible to commit to enforcing such an outcome, if needed, in 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 such punishment were needed, if that's the worst outcome for everyone). That is, suppose that outcome c does not belong to the acceptability correspondence at states W 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 behind this result is that if b and not a has to be selected at state H (as would be the case according to 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 to deter deviations, 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 W, 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 W 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 interests, 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).[8] The intuition for this result is that, in environments with a continuum and convex

[8] Yet, as we further show in another result, as long as Nash Equilibrium is maintained as solution concept, *Perfect*

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 is obviously characterized by important 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, there exist some economically important settings in which they can be guaranteed under standard and generally weak conditions.

In the next subsection we briefly discuss the most closely related literature. In Section 2 we introduce the model and baseline concepts. Section 3 introduces Safe Nash Implementation, with the results on the necessary and sufficient conditions in Sections 4 and 5, respectively. Section 6 contains the applications and some results for general solution concepts. Section 7 concludes.

1.1 Related Literature

The two closest papers are Eliaz (2002) and Shoukry (2019). Eliaz (2002) 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. Shoukry (2019) instead maintains Nash equilibrium as a solution concept, and 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. As we

Safety (i.e., ensuring that also deviations induce the same outcome as the SCF) can only be achieved for constant SCFs, regardless of the k parameter and of the restrictions on preferences.

note in the following (Corollary 1 below; see also Shoukry (2019)), this implies that the SCF is constant. To allow for more positive results, Shoukry (2019) allows for transfers and non-standard preferences, in which agents have a preference for truthful reporting. In contrast, in this paper we maintain standard preferences, we study general SCC (not only SCFs), and we allow a wider acceptability correspondence.

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 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* (see below).[9] 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.[10]

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. In a similar flavour, Tumennasan (2013) studies implementation under quantile

[9] Abreu and Sen (1991)'s virtual implementation also has a similar flavour. Within this framework the allocation rule are lotteries over sets of alternatives, and virtual implementation is achieved as long as the desired outcomes are implemented with probability at least $1 - \epsilon$. In this sense, in the space of lotteries, they require something almost perfectly safe' implementation (see below). However, no restrictions are imposed on the alternatives that are chosen with probability ϵ. Further to this, that notion does not address the concerns with robustness to deviations from the equilibria. Similarly, Abreu and Matsushima (1994) show that implementation in iterative elimination of weakly dominated strategies with the introduction of small fines, and hence with allocations that are close in the dimension of transfers.

[10] 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.

response equilibrium, letting the logit parameter approach the perfect rationality benchmark. In our paper, in contrast, we maintain Nash equilibrium and capture the possibility of mistakes as an extra desideratum, on top of the standard notion of implementation. Bochet and Tumennasan (2022) also maintain Nash Equilibrium as solution concept, 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), and Kneeland (2022).

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.

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*, $\mathcal{S}$, which for any given mechanism $\mathcal{M}$ induces a correspondence $\mathcal{S}^{\mathcal{M}} : \Theta \rightarrow 2^M$ 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 $g(\mathcal{S}^{\mathcal{M}}(\theta)) := \{x \in X : \exists m \in \mathcal{S}^{\mathcal{M}}(\theta) : g(m) = x\}$ denote the set of outcomes that are induced by action profiles that are consistent with the solution concept $\mathcal{S}$, at the state of the world θ. Full (strong) implementation is defined as follows:

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

For instance, if $\mathcal{S}$ is such that $\mathcal{S}^{\mathcal{M}}(\theta)$ denotes the set of Nash Equilibria of $G^{\mathcal{M}}(\theta)$ (i.e., $\mathcal{S}^{\mathcal{M}}(\theta) := \{m^* \in M : \forall i \in N, U_i^{\theta}(m^*) \geq U_i^{\theta}(m_i, m_{-i}^*)\}$), then the standard notion of *Nash Implementation* (Maskin, 1978) 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 $\mathcal{S}$-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*, $A : \Theta \rightarrow 2^X \setminus \emptyset$, 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 $A : \Theta \rightarrow 2^X \setminus \emptyset$ is *minimally safeguarding* if, for all $\theta \in \Theta$,

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

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, $W : X \times \Theta \rightarrow \mathbb{R}$, 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 utilitiarian (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 too, 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 $m^* \in \mathcal{S}^{\mathcal{M}}(\theta)$ 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 selected 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:[11]

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

First note that, for any $\mathcal{S}$, this notion generalizes the canonical notion of (non safe) implementation of Definition 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, this definition also generalizes existing notions in the literature, such as *outcome-robust implementation* (Shoukry, 2019) and *Fault Tolerant Implementation* (Eliaz, 2002), that share a similar motivation to ours.

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 $\mathcal{S}$-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.

Example 3 Consider again the environment in Ex.1: it will follow from our results that a SCF such that $f^*(W) = f^*(H) = a$ and $f^*(M) = b$ is Safe Implementable (letting the solution concept, $\mathcal{S}$, 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**.[12] 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).

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

[12] Shoukry (2019)'s *outcome-robust implementation* corresponds to this case, with $\mathcal{S}$ equal to Nash Equilibrium, only considering SCF, but allowing transfers and assuming that players have preferences for truthtelling.

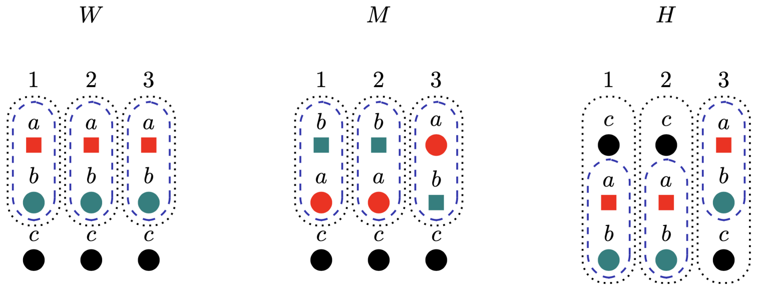

Figure 2: Firms 1, 2 and 3's preference orderings over the three alternatives, at the three states, W, M, and H. For each state, the allocation chosen by SCF f^* in Ex. 3 is indicated by a square. The acceptability correspondence A from Ex.1 is shown by the dotted lines, and is not maximally safe for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

Third, if the solution concept is held fixed across k (for instance, if $\mathcal{S}$ 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 $\mathcal{S}$ 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 $k' > k$.[13]

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 $\mathcal{S}^{\mathcal{M}} : \Theta \to 2^M$ 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 is such that Safe Nash Implementation is impossible for all sub-correspondences of A, 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, 1978), which obtains as a special case of (A, k)-Safe Nash Implementation when the extra safety requirement is moot (which

[13] 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 tightens the implementation requirement, is that k-FTNE depends on k in such a way that, if $\mathcal{M}$ is such that $\mathcal{S}_k^{\mathcal{M}}(\theta) \neq \emptyset \neq \mathcal{S}_{k'}^{\mathcal{M}}(\theta)$, and $k' < k$, then $\mathcal{S}_k^{\mathcal{M}}(\theta) \subseteq \mathcal{S}_{k'}^{\mathcal{M}}(\theta)$. The monotonicity in k that holds when $\mathcal{S}$ is held constant (as is the case for the main focus of our paper, in which $\mathcal{S}$ is standard Nash Equilibrium) is thus not guaranteed for FTI.

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 $k' \geq k$ – that is, increasing the number of deviations the mechanism must be resilient to makes implementation harder.

As it is well known, Maskin (1978) 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 (1978) also showed that, together with the following 'no veto condition', monotonicity is also a sufficient condition 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 $x \in F(\theta)$.*

In the next two sections we will 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. The sufficient conditions that we provide will also be a generalization of Definition 5, and we will show that they coincide with the necessary conditions in environments that satisfy a standard single-crossing property.

4 Necessity

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

Definition 6 (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 A admits a sub-correspondence A^* that satisfies the following conditions:*

1. *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. *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 $A^*(\theta) \subseteq A^*(\theta')$.*

If these conditions hold with $A^* = A$, then we say that A is tightly *w-Comonotonic.*

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. 6 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. 6 holds vacuously, letting $A^* = X$, and part 1 coincides with (Maskin) Monotonicity. However, if the A-correspondence entails non-trivial acceptability restrictions, then part 1 of Def. 6 restricts the SCC more than (Maskin) Monotonicity does. For a SCF, for instance, weak Comonotonicity 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 $L_i(f(\theta), \theta) \not\subseteq L_i(\theta, \theta')$. In the latter case, (Maskin) Monotonicity alone would leave the SCF free to set $f(\theta') \neq f(\theta)$, but weak Comonotonicity would not. More specifically, 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. In particular, if part 1 of Def. 6 is satisfied by a sub-correspondence of A, then it also holds for A.

The second part of Def. 6 states a monotonicity property of the A^* sub-correspondence, akin to Maskin's monotonicity for SCC, which imposes a lower bound on the inclusivity of the acceptability correspondence. 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. 6. 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, weak Comonotonicity gets less restrictive as the admissibility correspondence gets more inclusive: if A is weakly Comonotonic and $\hat{A}$ is such that $A(\theta) \subseteq \hat{A}(\theta)$ for all $\theta \in \Theta$, then also $\hat{A}$ is weakly comonotonic.

Example 4 Consider again the environment in Example 3 (see Fig.2). Both the A and A^* correspondences are (weakly) comonotonic with respect to the SCF f^* in that example: it is easy to check that A^* satisfies both conditions in Def. 6, and hence both A and A^* admit a sub-correspondence that satisfies those conditions. However, only A^* is tightly co-monotonic, since correspondence A in itself violates part 2 of Def. 6: moving from state $\theta = H$ to $\theta' = W$, allocation $a = f^*(H)$ moves (weakly) up in everyone's ranking within the set $A(H) = X$. Yet, $A(H) \not\subseteq A(W)$. This is obviously not the case for the A^* correspondence, since $A^*(H) = A^*(W) = \{a, b\}$. $\square$

We can now state our main result on necessity:

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

To understand the intuition behind this result, first consider the case in which the SCC is (A, k)-Safe Nash Implementable, and A is maximally safe. In that case, Theorem 1 implies that (A, F) are *tightly* w-Comonotonic, i.e. parts 1 and 2 of Def. 6 are satisfied with $A^* = A$. If 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.

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 $A(\theta)$. (If A is not maximally safe, the same logic applies, except not necessarily to A itself, but to some sub-correspondence of it.)

Example 5 In the usual environment from Ex. 3 (see Fig.2), for instance, the (weak) comonotonicity of (f^*, A) (which we discussed in Ex. 4) follows from Theorem 1 and from the fact that they are Safe Nash Implementable. As discussed in Ex.3, however, acceptability correspondence A is not *maximally safe* with respect to f^*, and hence the implication of Theorem 1 is that (f^*, A) are comonotonic, not necessarily tightly so. In fact, as discussed in Ex. 4, (f^*, A) are comonotonic, but not tightly, whereas (f^*, A^*) are *tightly comonotonic*. This tightness does follow from Theorem 1, because it is easy to verify that no sub-correspondence of A^* satisfies points 1 and 2 of Def.6, and hence A^* is maximally safe with respect to f^*. □

Theorem 1 formalizes a trade-off between the restrictiveness of the acceptability correspondence and the way that 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.6, 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.[14]

Theorem 1 also has the following direct and important implication.

[14] 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 13).

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. 6: 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 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), both of which 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 of Nash Equilibrium: since k-FTNE refines Nash Equilibrium (and more so, as k increases), it makes it easier to avoid 'bad' equilibria.[15] 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.

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 6 Consider an environment with two states, three alternatives, and for agents, denoted respectively as $\Theta = \{L, R\}$, $X = \{a, b, c\}$, and four agents: $N = \{1, 2, 3, 4\}$. 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 (formally, $L_3(b, L) \cap \{a, b\} = \{a, b\} \not\subseteq L_3(b, R) \cap \{a, b\} = \{b\}$). 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(formally, $L_1(c, L) \cap \{a, c\} = \{a, c\} \not\subseteq L_1(c, R) = \{c\}$.) Indeed, it will be easy

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

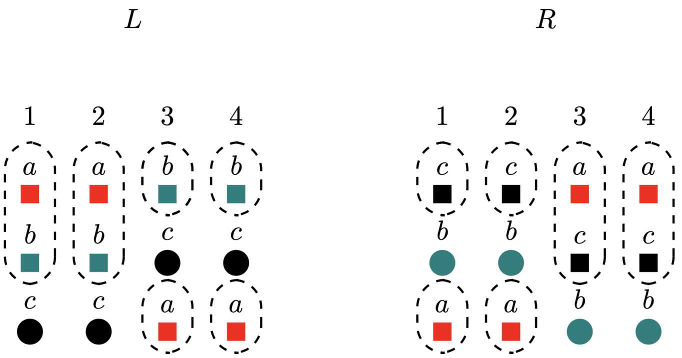

Figure 3: Players 1, 2, 3 and 4's preference orderings over the three alternatives, at the two states, L and R. For each state, the allocation chosen by SCC F in Ex. 6 is indicated by a square. The acceptability correspondence A is shown by the dashed lines, and is *perfectly safe*, as it coincides with the SCC at every state.

to verity 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 $R_k(\theta) = \bigcup_{m^* \in \mathcal{S}^{\mathcal{M}}(\theta)} B_k(m^*)$, where $\mathcal{S}^{\mathcal{M}}(\theta)$ 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 tightly w-Comonotonic sub-correspondence of A. If, moreover, A is maximally safe, then the implementing mechanism is k-surjective on A itself.*

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 *tightly w-Comonotonic* (i.e., if it satisfies conditions 1 and 2 in Def. 6) and a sub-correspondence of A. This directly implies that A and F are weakly Comonotonic, and hence Theorem 1 follows.

Next, 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 tightly w-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. 6), or to 'reduce' the dependence of the SCC on the state of the world (if A^k violates part 1 of Def. 6). 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 7 (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 A admits a sub-correspondence A^* that satisfies the following conditions:*

1. *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. *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 $A^*(\theta) \subseteq A^*(\theta')$.*

If these conditions hold with $A^* = A$, then we say that A is tightly *s-Comonotonic.*

First, notice that the difference between *Strong* and *Weak Comonotonicity* (Def. 6) 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. 7, 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 sufficiency result will show that, together with a generalization of Maskin's No-Veto condition, *tight s-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. 6 or 7) is both necessary and sufficient. We introduce next the notion of Safe No Veto:

Definition 8 (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 $\exists i \in N, \theta' \in \Theta : \forall j \in N \backslash \{i\}, x \in \text{argmax}_{y \in A(\theta')} u_j(y, \theta)$.*

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. 8 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, this condition has no bite if preferences rule out almost unanimity, as is the case in *economic environments*, in which in which agents have opposing interests over the allocations (e.g., Mirrlees (1976), Spence (1980), Arya et al. (2000)), such as the single-crossing environments that we will consider in Section 6. Furthermore, as we will explain below, Safe No-Veto is almost necessary. Nonetheless, it ensures a fairly strong Sufficiency result:

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

In the next subsection we provide a sketch of the proof, which as typical within the implementation literature is constructive. In Section 5.2 we further discuss the Safe No-Veto condition, some possible weakening, and we explain the sense in which it is almost necessary.

We also recall that, in the special case of SCFs, Theorems 1 and 3 directly imply the following:

Corollary 2 (Safe Implementation of SCF) *Let $f : \Theta \to X$ be such that (A, f) satisfy the Safe No-Veto condition (Def. 8). Then,f is (A, k) maximally Safe Nash implementable if and only if (A, f) are tightly* Comonotonic *(Def. 6 or 7, equivalently).*

5.1 Sketch of the Proof

As we mentioned, the proof is constructive in that it provides a mechanism that, under the conditions of the theorem, achieves Safe Nash Implementation. We present here the mechanism, and leave the remainder of the proof to the appendix. As it is standard within the implementation literature (see, e.g., Ben-Porath et al. (2019), Kartik and Tercieux (2012), Eliaz (2002), etc.), this mechanism shares the main structure of Maskin's *canonical mechanism*, in which each agent is asked to announce an outcome, a state, and a natural number.[16] Thus, $M_i = X \times \Theta \times \mathbb{N}$, with typical element $m_i = (x_i, \theta_i, n_i)$. Now, fix $k \in \mathbb{N} : 1 \leq k \leq \frac{n}{2}$, the outcome function is such that, for each $m \in M$, $g(m)$ is 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 $\exists j \in N : \forall i \in N \backslash \{j\}, m_i = (x, \theta, n^i)$ 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)* $\exists D \subset N$: $2 \leq |D| \leq k$, $m_i = (x, \theta, \cdot)$ and $x \in F(\theta)$ for all $i \in N \backslash D$, and $m_j \neq (x, \theta, \cdot)$ for all $j \in D$, then

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

  where $D^*(\theta, D) := \{j \in D : x_j \in A(\theta)\}$ and $i^*_D := \min\{i \in D^*(\theta, D) : n_i \geq n_j \quad \forall j \in D^*(\theta, D)\}$.

[16] Mechanisms of this kind, also known as *integer games*, are often adopted in the implementation literature in order to obtain the strongest sufficiency results. These mechanisms are inherently complex, and for this reason they have been criticized (see, e.g., Jackson (1992))). For recent work on full implementation via simpler and easier-to-interpret mechanisms, see, e.g. Ollár and Penta (2017, 2022a,b).

- *Rule (iv)* Otherwise, let $g(m) = x_{i^*}$, where $i^* := \min\{i \in N : n_i \geq n_j \quad \forall j \in N\}$.

In words, rule (i) ensures that if all agents agree report both the same state and an outcome, and the latter is within the SCC at said state, then this outcome is implemented by the mechanism. Rule (ii) acts as the typical "whistle-blowing" rule, which allows an agent to challenge an outcome by proposing an alternative, when all other agents are reporting the same state and outcome. The alternative proposed by the "whistle blower" is implemented only if two conditions are met: first, at the state reported by other players, the proposed alternative is not preferred by the agent over the outcome proposed by the others; second, if the proposed alternative is within the acceptability correspondence at the state reported by others. Rule (iii) sets up a standard integer game, to avoid that profiles in which agents' reports are not aligned form an equilibrium, with the difference of restricting the outcomes to be *within* the acceptability correspondence of the state that the majority of agents announce. This is a crucial difference compared to standard constructions, as it ensures that, even in the case of a large number of misreports, the outcomes are still deemed acceptable if most agents do not misreport, and it also ensures that the desirable equilibria fall within rule (i). This construction also ensures that, unless there all but one agents agree upon the top ranked alternative within the acceptability correspondence at the most commonly announced state, there cannot be any equilibria within this rule. Finally, rule (iv) uses a more standard integer game, which ensures that the only equilibria that can fall within this rule occur at states at which almost all players share the same top ranked alternative.

5.2 Safe No-Veto: Discussion and Extensions

As we briefly 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. 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.

Further to this, although the conditions are not too restrictive under most standard environments, as it is unusual to have preferences where almost all agents agree, it is possible to weaken the aspect of Safe No-Veto on the acceptability correspondence when agents have a weak preference for reporting correctly, in the following sense:

Definition 9 (Weak Preference 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 preference for correctness *in the mechanism $\mathcal{M}$ if $u_i : X \times \Theta \times M_i \to \mathbb{R}$ 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 $y \neq x$.*

The preference for correctness in this definition is *weak* in the sense that it is lexicographically subordinate to standard outcome-based preferences. Similar conditions for honesty have been studied, for instance, by Dutta and Sen (2012), Kartik et al. (2014), and Lombardi and Yoshihara (2020). This property essentially ensures that, keeping the outcome of the mechanism constant, agents would prefer to announce the correct state and allocation. This allows to weaken the notion of Safe No-Veto by dropping the condition that requires the acceptability correspondence to include all outcomes at state θ. Formally:

Definition 10 (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 $\exists i \in N, \theta' \in \Theta : \forall j \in N \backslash \{i\}, x \in \operatorname{argmax}_{y \in A(\theta')} u_j(y, \theta)$.*

To see why Safe No-Veto can be weakened in this way, notice that within the mechanism in Theorem 3 that condition plays two roles. First, similar to the standard No-Veto condition in Maskin (1978), it ensures that the equilibria at state θ that do not fall into rule (i) of the mechanism, are still consistent with the SCC at θ. Second, and unlike Maskin's case, Safe No-Veto plays the additional role of ensuring that said equilibria are themselves safe, in the sense that further deviations remain within the acceptability correspondence. It is the concern for safety of such 'extra' equilibria that causes the designer to concede that $A(\theta) = X$ when almost all agents agree. Such an extreme concession is required as there may be equilibria at θ that are exactly $\kappa < k$ deviations away from some equilibrium at θ'. Hence, an additional $k > k - \kappa > 0$ deviations, which would need to be considered for the safety of the equilibrium at θ, would lead to anything in X due to rule (iv).[17] Therefore this concession is needed. When there is a preference for correctness, however, the second role of Safe No-Veto is no longer needed, and hence the condition can be weakened in the sense above. This is because, although a deviation from such a message may not have been profitable when there was no preference for correctness, now the inclusion of preference for correctness allows for profitable deviations in the messages. This is because it allows for a deviation that does not change the allocation, and therefore not causing a first order change in the preferences, while providing a more correct message and therefore leading to a second order improvement. Given this, agents announcing the correct message, that would coincide with the equilibria at rule (i), is the only message where a profitable deviation does not exist. Thus, we

[17] Indeed, for this class of mechanisms, in which any allocations that is within the acceptability correspondence is reachable in strictly less than k deviations, and any other allocation is reachable in less that k steps from all others, this condition in necessary by this exact logic, when preferences only depend on the allocation.

conclude that, under weak preferences for correctness, the Nash equilibria only fall into rule (i), and therefore the acceptability correspondence need not satisfy more than tight s-Comonotonicity.

Proposition 1 *If $n \geq 3$, agents satisfy a weak preference for correctness, (F, A) are tightly s-Comonotonic and satisfy* weak *Safe No-Veto, then F is (A, k)-Safe Nash Implementable for all $k \in \mathbb{N} : 1 \leq k \leq \frac{n}{2}$.*

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 settings, 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. 11. 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 environment 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 $x_i = (x^1_i, x^2_i)$, with x^g_i 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 $\Theta_i = \{\theta^1_i, ..., \theta^{l_i}_i\} \subset \mathbb{R}_+$ that pin down i's preferences over X, labelled so that $\theta^1_i < ... < \theta^{l_i}_i$. 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 in both x^1_i and x^2_i for each $\theta_i \in \Theta_i$. Finally, we assume that preferences satisfy a standard single-crossing condition (SCC), that requires that agents' marginal rates of substitution between good 1 and good 2 is increasing in θ_i for each i.[18]

Letting $f : \Theta \to \mathbb{X}$ 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

[18] Formally, for any $x_i = (x^1_i, x^2_i)$ and $\theta_i, \theta'_i \in \Theta^i$ such that $\theta_i < \theta'_i$:

$$\frac{\frac{\partial u_i}{\partial x^2_i}(x, \theta_i)}{\frac{\partial u_i}{\partial x^1_i}(x, \theta_i)} < \frac{\frac{\partial u_i}{\partial x^2_i}(x, \theta'_i)}{\frac{\partial u_i}{\partial x^1_i}(x, \theta'_i)}.$$

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 SCC in this environment, 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* any SCF s.t. $f(\theta) \in int(X)$ for all $\theta \in \Theta$ 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 2 *Suppose that $n \geq 3$, and that the SCC 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 $k < \frac{n}{2}$.*

In fact, this result can be obtained with a mechanism that does not rely on the integer game. As we did above, we state the mechanism here, but reserve the proof for the appendix: Fix $\epsilon(\theta)$ such that $\mathcal{N}_{\epsilon(\theta)}(f(\theta)) \subseteq A(\theta)$ for all $\theta \in \Theta$. Let each agent $i \in N$ announce an allocation $x(i) \in X$ and a state $\theta(i) \in \Theta$. That is, $M_i = X \times \Theta$, with typical element $m_i = (x(i), \theta(i))$. The outcome function is defined so that $g(m)$ satisfies the following:

- *Rule (i)* If $\theta(i) = \theta$ for all $i \in N$, then $g(m) = f(\theta)$.
- *Rule (ii)* $\theta(i) = \theta$ for all $\in N \backslash \{j\}$ and $m_j = (x(j), \theta(j))$ for $\theta(j) \neq \theta$, then

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

- *Rule (iii)* If $\exists D \subset N$ agents such that $k \geq |D| > 1$ $m_1^i = \theta, \forall i \in N \backslash D$ then $g(m)$ is constructed as follows: $\forall i \in D$ let $\tilde{x}(i) = x(i)$ if $x(i) \in \mathcal{N}_{\frac{1}{|D|+1}\epsilon}(f(\theta))$ and $\tilde{x}^i = \lambda_i x(i) + (1 - \lambda_i) f(\theta)$ such that $d(f(\theta), \tilde{x}(i)) = \frac{1}{|D|+2}\epsilon$, $\lambda_i \in (0, 1)$ otherwise, where ϵ is fixed across agents such that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$. Now let $g(m) = f(\theta) + \sum_{i \in D} \tilde{x}(i)$.
- *Rule (iv)* Otherwise, let $g(m) = \frac{1}{n} \sum_{i \in N} x(i)$.

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.[19] Finally, we assume that the SCF and

[19] 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 acceptability correspondence are such that: (A.1) the SCF is efficient; (A.2) not assigning the object is always acceptable; and (A.3) whenever agent i is the designated winner, some other allocation is also acceptable.[20] Under these assumptions, the following possibility result obtains:

Proposition 3 *If $n \geq 3$ and preferences satisfy assumptions P.1-3, any (A, f) 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 standard (see, e.g., Eliaz 2002) and quite weak. Given the minimality of the 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.

As usual, we leave the proof to the appendix, but state the mechanism here: For each $i \in N$, let $M_i = X \times \mathbb{R}^n_+$, with a typical message $m_i = (j, v) \in X \times \mathbb{R}^n_+$, and define $g : M \to X$ as follows:

- *Rule (i)* If $m_i = (k, v)$ for all $i \in N$, with $v = \theta \in \Theta$ and $k = f(\theta)$, then $g(m) = k = f(\theta)$
- *Rule (ii)* If $m_i = (k, v)$ for all $i \in N \backslash \{j\}$, with $v = \theta \in \Theta$ and $k = f(\theta)$ and $m_j = (l, \cdot)$, $l \neq k$, then
$$g(m) = \begin{cases} l & \text{if } l \in L_j(k, \theta) \cap A(\theta) \\ k & \text{if } l \notin L_j(k, \theta) \cap A(\theta) \end{cases}$$
- *Rule (iii)* If $m_i = (k, v)$ for all $i \in N \backslash D$, with $2 \leq |D| < \frac{n}{2}$, $v = \theta \in \Theta$ and $k = f(v)$, and for all $j \in D$, $m_j \neq (k, \cdot)$ then
$$g(m) = \begin{cases} l^{i^*} & \text{if } D^*(\theta, D) \neq \emptyset \\ k & \text{if } D^*(\theta, D) = \emptyset \end{cases}$$
where $D^*(\theta, D) = \{j \in D | l^i \in A(\theta)\}$ and $i^* = \min\{i \in D^*(\theta, D) | v^i_i \geq v^j_j \quad j \in D^*(\theta, D)\}$.
- *Rule (iv)* otherwise let $g(m) = l^{i*}$ where $i^* = \min\{i \in N | v^i_i \geq v^j_j \quad j \in N\}$

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 11 *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 $\mathcal{S}$ (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:

[20] 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, $\exists x \neq i, \emptyset$ s.t. $x \in A(\theta)$ whenever $f(\theta) = i$.

Proposition 4 *Suppose that Θ is rich, $1 < |X| \leq n$. No SCF is (A,k)-Safe $\mathcal{S}$-Implementable for some $k > 1$, 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 5 *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 where $x = f(\theta')$, and clearly $x \neq z = f(\theta)$. By richness, there is a state θ'' where 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 the SCF, this is a contradiction.

7 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 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 deviates 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 considerations, 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 (see, 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'. 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 agents must 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 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, 1978). While we also offer some results for general 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 (cf. Eliaz (2002); Shoukry (2019)); (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, 2022a,b), etc.), is yet another direction for future research.

Appendix

A Proofs

Proof of Theorem 1: Suppose that F is (A,k)-Safe Nash Implementable. Therefore there is some mechanism $\mathcal{M}$ that (A,k)-Safe Implements F. For such a mechanism, take A^* to be such that $A^*(\theta) = g(\{m \in M | d(m,m^*) \leq k, \quad m^* \in \mathcal{S}^{\mathcal{M}}(\theta)\})$. Note that by definition we have that $A^*(\theta) \subseteq A(\theta)$ and therefore A^* is a sub-correspondence. By definition, if A is maximally safe then

$A^*(\theta) = A(\theta)$.

We will show that F and A are comonotonic, using the sub-correspondence A^*, 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^* 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 $g(m_i', m_{-i}^*) \in g(\{m \in M | d(m, m^*) \leq k, \quad m^* \in \mathcal{S}^{\mathcal{M}}(\theta)\}) = A^*(\theta)$. 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 $L_i(x,\theta) \cap A^*(\theta) \subseteq L_i(x,\theta') \cap A^*(\theta)$.

Now we show that for $\theta, \theta' \in \Theta$, 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 $x \in F(\theta)$ and $L_i(x,\theta) \cap A^*(\theta) \subseteq L_i(x,\theta')$ $\forall i \in N$. 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. $u_i(g(m_i', m_{-i}^*), \theta') > u_i(f(\theta), \theta')$. We conclude that $g(m_i', m_{-i}^*) \in X \backslash L_i(f(\theta), \theta')$. By (A,k)-Safe Nash Implementation, and the definition of $A^*(\theta) = g(\{m \in M | d(m, m^*) \leq k, \quad m^* \in \mathcal{S}^{\mathcal{M}}(\theta)\})$, it must be that $g(m_i', m_{-i}^*) \in L_i(f(\theta), \theta) \cap A^*(\theta)$. A contradiction to $L_i(f(\theta), \theta) \cap A^*(\theta) \subseteq L_i(f(\theta), \theta')$.

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 a k agents misreporting from Equilibrium at θ are also reached within k deviations of an Equilibrium at θ'. Concluding that $A^*(\theta) = g(\{m \in M | d(m, m^*) \leq k, \quad m^* \in \mathcal{S}^{\mathcal{M}}(\theta)\}) \subseteq A^*(\theta')$. By definition we have that $A^*(\theta) \subseteq A^*(\theta')$.

Given that if By definition, if A we maximally safe then $A^*(\theta) = A(\theta)$, we conclude that if a social choice function is maximally safe it must be tightly w-comonotonic. ■

Proof of Theorem 2:

The result follows directly from the construction used in the proof of theorem 1. ■

Proof of Theorem 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 $m_j = (y, \cdot, \cdot)$ 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 a 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. In order 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), a contradiction to $u_j(y, \theta) > u_j(x, \theta)$. Therefore it must be that m^* is a Nash Equilibrium leading to $x \in F(\theta)$.

Step 2. We want to show that $\nexists m^*$ such that m^* is a Nash Equilibrium at θ such that $g(m^*) \notin F(\theta)$. To do so we proceed by splitting into 4 cases, corresponding with the 4 sections of the mechanism.

Case 1: Suppose that m^* is such that $m_i^* = (x, \theta', \cdot)$, $\forall i \in N$ and $x \in F(\theta')$. Suppose that the true state is $\theta \neq \theta'$ where $x \notin F(\theta)$ and m^* is a Nash Equilibrium.

By rule (ii) the set of attainable alternatives for all $i \in N$ is $L_i(x, \theta') \cap A(\theta')$ which, as m^* is a Nash Equilibrium at θ, implies that $\forall y \in L_i(x, \theta') \cap A(\theta')$, $u_i(x, \theta) \geq u_i(y, \theta)$, and therefore $y \in L_i(x, \theta) \cap A(\theta'), \forall y \in L_i(x, \theta') \cap A(\theta'), \forall i \in N$. Therefore $L_i(x, \theta') \cap A(\theta') \subseteq L_i(x, \theta) \cap A(\theta'), \forall i \in N$ and therefore by tight s-Comonotonicity $x \in F(\theta)$ and $A(\theta') \subseteq A(\theta)$. Therefore as deviations lead to allocations in $A(\theta')$ they are still considered (A, k)-safe.

Case 2): Suppose that m^* is a Nash Equilibrium at θ and is of the form $m_i^* = (x, \theta', \cdot), \quad \forall i \in N \backslash \{j\}$, $m_j^* = (y, \cdot, \cdot)$.

If $k = 1$ the set of attainable alternatives $\forall i \in N \backslash \{j\}$ is X and therefore implies that $g(m^*) \in \text{argmax}_{z \in X} u_i(z, \theta), \quad \forall i \in N \backslash \{j\}$ in order for m^* to be a Nash Equilibrium. By safe no-veto, $g(m^*) = x \in F(\theta)$ and $A(\theta) = X$, and therefore as deviations lead to rule iv) where X is attainable, they are still considered safe.

If $k > 1$, the set of attainable alternatives $\forall i \in N \backslash \{j\}$ is $A(\theta')$ via rule (iii). This implies that $g(m^*) \in \text{argmax}_{z \in A(\theta')} u_i(z, \theta), \quad \forall i \in N \backslash \{j\}$. Safe no-veto implies that $g(m^*) = x \in F(\theta)$ and $A(\theta) = X$, and therefore any deviation of k agents from the Equilibrium m^* can only lead to rule

iii) or iv) where at most the allocations X are available, and therefore is considered safe.

Case 3: Suppose that m^* is a Nash Equilibrium at θ such that $m_i^* = (x, \theta', \cdot)$, $\forall i \in N \backslash D$ where $2 \leq |D| \leq k$, $x \in F(\theta')$ and $m_j^* \neq (x, \theta', \cdot)$, $\forall j \in D$.

$\forall i \in N$ the set of attainable alternatives options is at least $A(\theta')$ (via rule iii) or X via rule iv)). It follows by m^* being a Nash Equilibrium at θ that $g(m^*) \in \text{argmax}_{z \in A(\theta')} u_i(z, \theta), \quad \forall i \in N$. By safe no-veto this implies that $g(m^*) \in F(\theta)$. It also implies that $A(\theta) = X$, and therefore any deviation from this Equilibrium leading to either rule iii) or rule iv) of the mechanism is considered safe.

Case 4: Suppose that m^* falls into none of the previous cases and is a Nash Equilibrium at θ. If $k = 1$ then by safe no-veto $g(m^*) \in \text{argmax}_{z \in X} u_i(z, \theta), \quad \forall i \in N$ implies that $g(m^*) \in F(\theta)$ and $A(\theta) = X$ and therefore any deviation is considered safe. If $k > 1$ then the set of attainable alternatives for all $i \in N$ is at least $A(\theta')$ where θ' is the most commonly announced state. It follows that $\exists j \in N$ such that for state θ' $g(m^*) \in \text{argmax}_{z \in A(\theta')} u_i(z, \theta)$ $\forall i \in N \backslash \{j\}$. This implies that $g(m^*) \in F(\theta)$. Further, $A(\theta) = X$, and therefore any deviation from m^* is considered safe.

Step 3: Now we must show that $\forall m^*$ such that m^* is a Nash Equilibrium at θ, $\forall D_k$, $\forall m_{D_k} \in \times_{i \in D_k} M_i$, $g(m^*_{-D_k}, m_{D_k}) \in A(\theta)$. This follows by construction of the mechanism and the examination of all possible Equilibria in the previous step. ■

Proof of lemma 1:

Suppose for some $m^{\star,\theta}$ we have that $x \in g(B_{k-1}(m^{\star,\theta})) \cap \text{argmax}_{y \in A(\theta)} u_i(y, \theta')$ $\forall i \in N$. 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 1:

Take the same mechanism and logic as proposition 3. Now all is left to show is that all Equilibria are contained in case 1 whenever $k \geq 1$.

Suppose that we have an Equilibrium in case (ii). Suppose the state is θ'. It must be that $g(m^*) \in F(\theta)$ while some at least some agents disagree on $x = g(m^*)$ or θ. For this to be Equilibrium, it must be that no other agent but the one report something other than x and θ wishes to deviate to any other message. We therefore conclude that x is in arg max of either

1. X in the case that $k = 1$, and therefore it must be that $x \in \text{argmax}_{y \in A(\theta'')} u_i(y, \theta')$ for some θ''

2. or $A(\theta)$ when $k > 1$

when the state is θ'. $\theta' = \theta$, clearly the agent who is reporting other than θ and x can be provided with ϵ more utility for correctly reporting, therefore this message cannot be a Nash Equilibrium. If $\theta' \neq \theta$ then by weak Safe no-veto we conclude that $x \in F(\theta')$. However, then all agents who report θ can increase their utility by ϵ by reporting θ', concluding that this message cannot be a Nash Equilibrium.

Suppose that the Equilibrium is in case 3 or 4. Here it must be that all agents agree that $g(m^*) \in \text{argmax}_{z \in A(\theta')} u_i(z, \theta)$. However, the set of attainable outcomes is at least $A(\theta')$, where $g(m^*) \in A(\theta)$, so for any agent not announcing $x = g(m^*)$ and the true state, they can increase their utility by announcing this. ■

Proof of lemma 2:

Take $\theta, \theta' \in \Theta$ such that $f(\theta) = x \neq f(\theta')$. Let agent i be such that $\theta^i \neq \theta'^i$. Let $\bar{u}_i = u_i(f(\theta), \theta^i)$ and $\bar{u}'_i = u_i(f(\theta), \theta'^i)$. 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 2:

Let each agent $i \in N$ announce an outcome and the state. Therefore $M_i = X \times \Theta$, with typical element $m_i = (x^i, m_1^i)$ Let $g(m)$ be as follows:

Rule (i) If $m_1^i = \theta \quad \forall i \in N$ then $g(m) = f(\theta)$.

Rule (ii) $m_1^i = \theta \quad \forall i \in N \backslash \{j\}$ where $m_j = (x^j, \theta')$

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

Rule (iii) If $\exists D \subset N$ agents such that $\alpha \geq |D| > 1$ $m_1^i = \theta, \forall i \in N \backslash D$ then $g(m)$ is constructed by the following $\forall i \in D$ let $\tilde{x}^i = x^i$ if $x^i \in \mathcal{N}_{\frac{1}{|D|+1}\epsilon}(f(\theta))$ and $\tilde{x}^i = \lambda^i x^i + (1 - \lambda^i) f(\theta)$ such that $d(f(\theta), \tilde{x}^i) = \frac{1}{|D|+2}\epsilon$, $\lambda^i \in (0, 1)$ otherwise, where ϵ is fixed across agents such that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$. Now let $g(m) = f(\theta) + \sum_{i \in D} \tilde{x}^i$.

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 $\theta' \neq \theta$. In turn, this implies that $\exists i$ such that $\theta^i \neq \theta'^i$. For this agent $\exists x^i$ 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 an Equilibrium.

Case 2: It is clear that in rule (ii) we can not have any Equilibria, any agent j who is announcing $m_1^j \neq \theta$ and $m_1^i = \theta$ $\forall i \in N\backslash\{j\}$ can monotonically increase their allocation by announcing an allocation to the north east of the current allocation, say x'^j, while $x'^j \in L_j(f(\theta),\theta)\cap\mathcal{N}_{\frac{\epsilon}{2}}(f(\theta))$, which can always be achieved by the open nature $\mathcal{N}_{\frac{\epsilon}{2}}(f(\theta))$ and continuity of $u_j(\cdot,\theta)$ (implying $X\backslash L_j(f(\theta),\theta)$ is open), and $\theta' \neq \theta$.

Case 3: By the same logic, there cannot be an Equilibrium in Rule (iii), any agent $i \in N\backslash D$ can announce an allocation to the north east of $\tilde{x}^i$ such that $x^i \in \mathcal{N}_{\frac{1}{|D|+1}\epsilon}(f(\theta))$, leading to rule (iii) or (iv), regardless, monotonically increase their allocation (notice this is the case due to the penalty for announcing $x^i \notin \mathcal{N}_{\frac{1}{|D|+1}\epsilon}(f(\theta))$ 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 as each agent 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.

Step 3: Here, the only Equilibria can lie in Rule (i). Deviations of size α or less all lead to $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$. ■

Proof of proposition 3:

Let $X = N \cup \{0\}$, where 0 represents the good being unallocated.

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

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

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

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

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

where

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

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

Rule (iv) otherwise let $g(m) = l^{i*}$ where $i^* = \min\{i \in N | v_i^i \geq v_j^j \quad j \in N\}$.

Clearly all messages that messages m^* at state θ that fall into rule i) are a NE. Now to show is that all NE are Safe. We will do so by showing that rule i) constitute the only NE.

Suppose that there is a Nash Equilibrium in rule ii). Suppose that $k = f(\theta)$ is the agent announcing $m_j \neq (k, v)$, therefore $g(m) = k$ or $g(m) = l$. Then there must be some other agent

$i' \in A(\theta)$, $i' \neq k$ who can announce $m_{i'} = (i', v') \neq (k, v)$ and $v'_{i'} > v^k_k$ ensuring they will be allocated the good, and therefore cannot be Equilibrium. Now suppose that some agent other than k announces $m_j \neq (k, v)$. Then $\exists i' \in A(\theta) \backslash \{j\}$ who can profitably deviate by announcing $m_{i'} = (i', v')$ with $v'_{i'} > v^j_j$.

Clearly there can be no Equilibria in rule *iii*) and *iv*).

Finally, suppose that there is some NE in rule i) m^* at θ such that, for some θ' such that $g(m^*) = f(\theta') = k \neq f(\theta)$. This cannot be the case as reverting to the empty allocation is attainable and by assumption gives a higher payoff than an undeserving agent. ■

References

Abreu, D. and Matsushima, H. (1994). Exact implementation. *Journal of Economic Theory*, 64(1):1–19.

Abreu, D. and Sen, A. (1991). Virtual implementation in nash equilibrium. *Econometrica: Journal of the Econometric Society*, pages 997–1021.

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

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

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

Bochet, O. and Tumennasan, N. (2022). One truth and a thousand lies: Focal points in mechanism design. *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.

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. (1978). Implementation and strong nash equilibrium.

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

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. (2022a). 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. (2022b). A network solution to robust implementation: The case of identical but unknown distributions.

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. 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.