Northwestern | Kellogg

CMS-EMS
Center for Mathematical Studies in
Economics and Management Sciences

Discussion Paper #1606

BEYOND DOMINANCE AND NASH: RANKING EQUILIBRIA BY CRITICAL MASS

(Forthcoming in Games and Economic Behavior)

Adam Tauman
OpenAI

Ehud Kalai
Northwestern University

December 29, 2023

J.L. Kellogg School of Management at Northwestern University
2211 Campus Drive, Evanston, Illinois 60208 | 847.491.3300 | kellogg.northwestern.edu

BEYOND DOMINANCE AND NASH: RANKING EQUILIBRIA BY CRITICAL MASS

ADAM TAUMAN KALAI AND EHUD KALAI

Abstract. Strategic interactions pose central issues that are not adequately explained by the traditional concepts of dominant strategy equilibrium (DSE), Nash equilibrium (NE), and their refinements. A comprehensive analysis of equilibrium concepts within the von Neumann-Nash framework of n-person optimization reveals a decreasing hierarchy of n nested concepts ranging from DSE to NE. These concepts are defined by the "critical mass," the number of players needed to adopt and sustain the play of a strategy profile as an equilibrium. In games with $n > 2$ players, the $n - 2$ intermediate concepts explain strategic issues in large social systems, implementation, decentralization, as well as replication studied in economics, operations management, and political games.

1. Introduction

For over a half a century, game theory has provided a framework for understanding strategic interaction across a wide range of disciplines. However, despite the popularity of dominant strategy equilibria (DSE) and Nash equilibria (NE) as tools for analyzing n-player games, they do not always capture the full range of strategic issues that arise in practice. While DSE are often touted as the most reliable form of equilibrium behavior, their nonexistence in many applications has led to the widespread use of the weaker notion of NE. In this paper, we introduce an extended framework

Date: Dec. 29, 2023.
2000 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18.
Key words and phrases. Nash equilibrium, Dominant strategy equilibrium, faulty players, social contract.
This paper includes results described in two earlier unpublished papers (Kalai, 2019; Kalai and Kalai, 2022).

of critical mass equilibria that bridges the gap between DSE and NE, and show how it enhances the strategic analysis of various games. By eliminating shortcomings associated with DSE and NE, critical mass equilibria offer a more robust and accurate description of strategic behavior with important implications for social, biological, and computational sciences.

The main theorem in this paper characterizes all the strategic equilibrium concepts in the von Neumann Nash (vNN) framework of n-player games ($n \geq 2$) and reveals the hierarchy of equilibrium concepts:

$$\{\text{DSE}\} = \mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \cdots \subseteq \mathcal{C}_n = \{\text{NE}\}.$$

The equilibrium concepts in this hierarchy are defined and arranged according to a well-defined *critical mass index* κ that specifies the number of players needed for the adoption of a profile of strategies as equilibrium play. The concept $\mathcal{C}_m$, which we refer to as *equilibrium of critical mass* m, consists of all strategy profiles of critical mass m or less: $\mathcal{C}_m = \{\pi \mid \text{profiles of strategies with } \kappa(\pi) \leq m\}$. Equivalently, $\kappa(\pi) = m$ if and only if $\pi \in \mathcal{C}_m \setminus \mathcal{C}_{m-1}$.

Generalizing the notion of incentive compatibility, one could say that $\mathcal{C}_m$ consists of the strategy profiles π that are *m-incentive-compatible* in that the mutual play of π by any group of m or more players is individually optimal, regardless of the strategies played by the group outsiders.

The closely related *index of resilience* against defections ρ, defined by $\rho(\pi) \equiv n - \kappa(\pi)$, is dual to the critical mass κ. Any group of $\rho(\pi)$ or fewer defectors cannot disrupt the best response properties of π: First, every member of such a group can only lose in any group defection. Moreover, if such a group still chooses to defect, they cannot incentivize any group outsider to defect.

1.1. **Earlier literature.** It is not surprising that the equilibrium considerations above have been the subject of discussion in earlier literature. Going back to stag-hunt games in the 1700s (prior to the birth of modern game theory), the philosophers Jean-Jacques Rousseau and David Hume studied the stability issues discussed above in the context of social decision making (see Skyrms, 2001). Schelling (1973) discusses an n-player prisoners' dilemma, centered on equilibria that are stable according to the resilience index ρ. Computer scientists' concerns about faulty computation in distributed computing, see Goldreich, Goldwasser, and Linail (1998), were important in motivating economists to study implementation that is robust against defection, see Eliaz (2002) and Abraham et al. (2006) below.

Eliaz (2002) had the goal of implementing socially efficient outcomes in environments in which some of the economic agents are subject to faulty equilibrium behavior. He showed that if an equilibrium π of an implementation game is *k-fault-tolerant* with $k \leq$ the number of faulty players, then such robust implementation is possible. It is easy to see that an equilibrium π is k-fault-tolerant in the sense of Eliaz iff $k \leq \rho(\pi)$. Similar robust implementations for problems of distributed computing and guessing games in computer science were presented by Abraham et al. (2006). Additional results on fault-tolerant equilibria in large games are presented in Gradwohl and Reingold (2014); and Deepanshu and Berry (2020) study sufficient conditions for a Nash equilibria to have low critical mass. In recent preliminary work, Kim, Min, and Wooders (2022) report on experimental results that show that in stag-hunt games players are significantly more likely to play equilibria π with lower critical mass values, $\kappa(\pi)$.

1.2. **Illustrative examples.**

Example 1. ***Rebellion tipping point***: *On a certain day, simultaneously, each member of a group of n=100,000 citizens must choose one of two strategies: rebel*

(R) or acquiesce (A). The known government policy is to randomly choose and jail for a day one rebel. We assume that for every citizen i the payoff of A is 0, no matter what the other citizens choose. However, if i chooses R, then the net payoffs are 2 if i is not jailed, and -1 if i is jailed. We consider two strategy profiles: full acquiescence $\bar{A}$, in which everybody acquiesces, and total rebellion $\bar{R}$, in which everybody rebels.

Clearly, a lone rebel is sure to be jailed and suffer the payoff of -1, and a rebel has a positive expected payoff (≥ 0.5) iff the number of rebels ≥ 2. This means that both $\bar{A}, \bar{R}$ are NE, and there are no DSE, i.e., $\bar{A}, \bar{R} \in \mathcal{C}_n$, and $\mathcal{C}_1 = \emptyset$. However, the intermediate critical-mass concepts reveal a significant difference between $\bar{A}$ and $\bar{R}$.

The critical mass needed to justify the total rebellion is low: $\kappa(\bar{R}) = 2$. This means that conditionally on *any* one citizen rebelling, it is a dominant strategy for every other citizen to rebel. For this reason we informally think of $\bar{R}$ as nearly dominant. This exact reasoning justifies our more general view: in any n-person game, any profile π with $\kappa(\pi) = 2$ is a *nearly dominant strategy equilibrium*.

On the other hand, a citizen prefers to acquiesce only if *all* the others do; thus, the critical mass needed for full acquiescence is high: $\kappa(\bar{A}) = n$. Therefore, the resilience-against-defection index is minimal, $\rho(\bar{A}) = n - \kappa(\bar{A}) = 0$, which means that one single defector (a rebel) is enough to motivate others to defect. In this sense, the full acquiescence profile $\bar{A}$ is fragile and, more generally, in any n-player game we call any profile π with resilience $\rho(\pi) = 0$ a *fragile Nash equilibrium.*

We note that despite the drastic difference in adoptability and sustainability of $\bar{R}$ versus $\bar{A}$, the standard equilibrium concepts of game theory treat them equally: both are Nash equilibrium and neither is a dominant strategy equilibrium.

Despite the simplicity of the analysis above, the low critical mass of full rebellion suggests that tipping into full rebellion may be accomplished by just a few rebels. To raise the tipping point, the government may consider a tougher punishment policy; for example it could arrest up to 100 random rebels, and jail each arrested rebel for one year. Assume for simplicity that the net payoffs under the tougher policy are as above, except that the net payoff of an arrested rebel is -365. It is easy to compute that the new expected payoff of any one of r (≥ 100) rebels is $-365 \times 100/r + 2(r - 100)/r$, which is non-negative iff $r \geq 18,350$; thus, the critical mass needed for full rebellion is $\kappa(\bar{R}) = 18,350$ rebels. This suggests that under the tougher policy tipping into full rebellion requires many thousands of rebels.

Example 2. ***Centralized Chip Production**: Player* 1 *is a chip producer, and players* $2, 3, \ldots, n$*, are chip users. Simultaneously, each of the* n *players must choose one of two types of chips, hard (*H*) or soft (*S*). The producer strictly prefers to produce* H*, i.e.,* $u_1(\theta) = 1$ *if* $\theta_1 = H$*, and* $u_1(\theta) = -1$ *if* $\theta_1 = S$*. Each user* i *strictly prefers to match the producer's choice, i.e.,* $u_i(\theta) = 1$ *if* $\theta_i = \theta_1$ *but* $u_i(\theta) = -1$ *at all other strategy profiles* θ*. Consider the profile* $\bar{H}$*, in which all the players choose* H*.*

Notice that $\mathcal{C}_n = \{\bar{H}\}$ is the unique Nash equilibrium of this game; in fact, it is the only profile that survives the sequential elimination of strictly dominated strategies. It also passes the traditional refinement tests: it is strong (a la Aumann, 1959), perfect (a la Selten, 1975), proper (a la Myerson, 1978), coalition-proof (a la Bernheim, Peleg and Whinston, 1987) and stochastically stable (a la Young, 1993; Kandori, Mailath, and Rob, 1993).

Informally, however, at this equilibrium the chip users are vulnerable in a disturbing sense: for instance, if due to unforeseen political events or acts of nature (not modeled in the game), the producer chooses S instead of H, each user is stuck with the wrong choice, H.

This vulnerability of the equilibrium $\bar{H}$ is detected formally by the critical mass analysis: a defection from H to S by a single player (the producer) is enough to motivate other players (chip users) to defect from H. This means that $\bar{H}$ has minimal resilience $\rho(\bar{H}) = 0$ ($\kappa(\bar{H}) = n$) and it is a fragile NE in our terminology.

Strategic analysis of tipping points, decentralization, and other topics discussed in the body of this paper have been studied by researchers in the social sciences, operations management, computer science, and other areas. It is important to note that the critical mass analysis unifies these discussions by addressing them through the critical mass of the equilibrium of a game, without alluding to further game modifications. The index of critical mass provides a single parameter that assesses the reliability of the assumed equilibrium behavior in these various applications.

As the examples above illustrate, our critical mass concepts provide a significant augmentation for game-theoretic analysis. For example, a large critical mass may serve as a valuable red flag to researchers who assume that a Nash equilibrium π presents a reliable description of the outcome of the strategic interaction they study, because it indicates low resilience to defection. On the other hand, analysts who use equilibria π with low critical mass, i.e., high resilience, may be correctly reassured that their equilibria present reliable descriptions of the possible outcome of the strategic interactions they study.

1.3. **Paper outline.** Section 2 focuses on the basic definitions and properties of critical mass index and equilibrium. It discusses a relationship of the resilience index to Eliaz (2002) work on implementation. In addition, this section presents the definition and basic properties of stag-hunt games that are used in the proof of the main theorem.

Section 3 presents the formal statement and proof of the main theorem of the paper. It presents a formal definition of the vNN framework used in the main theorem, including three minimal axioms that are satisfied by DSE and NE.

Section 4 focuses on additional examples and illustrations of the use of the critical mass concepts. These include a discussion of the role of the two indices in explaining equilibria observed in large social systems, equilibrium implementation, and equilibria in graph-matching games with implications for issues of decentralization and replications in operations management and political games.

Section 5 summarizes the contributions of this paper and discusses future work on alternative notions of critical mass indices within and outside the vNN framework.

2. Critical mass index and equilibrium

2.1. **Definitions and basic properties of critical mass.** The strategic games studied in this paper are defined for a set of n *players* $N = \{1, 2, \dots, n\}$ with $n \geq 2$. An n-player game is a *joint payoff function* $u : \Theta \to \mathbb{R}^n$, in which the domain of u has a product structure $\Theta = \times_{i \in N} \Theta_i$, with each Θ_i denoting the set of individual *strategies* of player i. Elements of Θ are referred to as the strategy *profiles* of u, and $u_i(\theta) \in \mathbb{R}$ is the *payoff to player* i when the profile $\theta \in \Theta$ is played. For brevity and without loss of generality, we may restrict the description of a game to its joint payoff function u alone, with the understanding that the set of profiles of the game is defined by $\Theta(u) \equiv \text{domain}(u)$, and that the set of strategies of player i is denoted by $\Theta_i(u) = \Theta_i(u_i)$. We assume that the sets of strategies Θ_i include mixed (randomized) strategies, if available, which removes the notational burden of separately defining mixed strategies. To formally define the set of all possible games, we consider the set of n-player games on strategies from a given superset $\overline{\Theta}$:

$$\mathcal{U}_n(\overline{\Theta}) \equiv \{u : \Theta \to \mathbb{R}^n \mid \Theta = \times_{i \in N} \Theta_i \text{ and } \Theta_1, \Theta_2, \dots \Theta_n \subseteq \overline{\Theta}\}.$$

We assume that $\overline{\Theta}$ is fixed and has at least $|\overline{\Theta}| \geq 3$ strategies (in fact, it would typically be infinite). We write $\mathcal{U}_n$ for $\mathcal{U}_n(\overline{\Theta})$.

Given a profile π, we say that player i is a π-*player* at a profile θ if $\theta_i = \pi_i$; and if $\theta_i \neq \pi_i$, we say that player i is a π-*defector*. When we have no information about θ_i, we may refer to player i as a potential defector.

Extending standard game theory conventions, for two profiles α, $\pi \in \Theta(u)$ and a subset of player $S \subseteq N$, we denote by (α_S, π) the profile in which all the players $i \in S$ play their α_i strategies and all the players $j \notin S$ play their π_j strategies. For any game u, a group of players $S \subseteq N$ and a profile $\pi \in \Theta(u)$, the *game played by* S *under* π, u_S^π, is described as follows: The set of players is S, the strategies of every player $i \in S$ are the same as their strategies in u (i.e., $\Theta_i(u_i)$). When the players i in S play a strategy profile α their payoffs are $u_i(\alpha_S, \pi)$. The singleton coalition $\{i\}$ may be simply denoted by i.

In any game u we say that π_i is a (weak) *best response* of player i to a profile θ if $u_i(\pi_i, \theta_{-i}) \geq u_i(x_i, \theta_{-i})$ for all $x_i \in \Theta_i(u_i)$. A profile π is best response to a profile θ if each π_i is best response to θ.

An *equilibrium concept* is a correspondence $\mathcal{E}$ that assigns to every game u a subset of its profiles, $\mathcal{E}(u) \subseteq \Theta(u)$. Elements of $\mathcal{E}(u)$ are referred to as $\mathcal{E}$-equilibrium of u, but when it is clear from the context, we may omit the specification of u. Two familiar examples are the Nash equilibrium, $\mathcal{E}(u) = NE(u) \equiv \{\pi \in \Theta(u) |$ each π_i is best response to $\pi\}$, and the dominant strategy equilibrium, $\mathcal{E}(u) = DSE(u) \equiv \{\pi \in \Theta(u)|$ each π_i is best response to any profile $\theta \in \Theta(u)\}$. There are two *trivial* equilibrium concepts: the concept $\mathcal{E}(u) \equiv \emptyset$ for all games u; and the concept $\mathcal{E}(u) \equiv \Theta(u)$ for all games u.

For any two profiles of individual strategies, θ and π, we define the agreement level of the pair by

$$a(\theta, \pi) \equiv |\{i \in n \mid \theta_i = \pi_i\}|. \tag{2.1}$$

The ball of $\geq k$ (at least k) agreements around a profile π is defined by

$$A_k(\pi) \equiv \{\theta \in \Theta(u) \mid a(\theta, \pi) \geq k\}. \tag{2.2}$$

Thus, $A_k(\pi)$ is the set of profiles with at least k *π-players.* A useful *dual* perspective is that $A_k(\pi)$ is the set of profiles that allow for at most $n-k$ *potential π-defectors*, i.e., a Hamming ball of radius $n-k$. Similarly, for strategy profiles of player i's opponents we use $A_k(\pi_{-i})$ to denote the opponents' profiles θ_{-i} that agree with π for at least k opponents: $A_k(\pi_{-i}) = \{\theta_{-i} \in \Theta_{-i}(u) \mid a(\theta_{-i}, \pi_{-i}) \geq k\}$.

Next, we formally introduce the critical-mass concepts.

Definition 1 (m-Incentive-Compatible)**.** *For any integer $1 \leq m \leq n+1$, a profile π is (uniformly) $m-$*incentive-compatible *(m-IC for short) in a game u, if at every profile θ with at least m π-players, π_i is a u-best response to θ for each π-player.*

The term *uniformly* emphasizes that when π is m-IC, it is incentive compatible for *all* strategy profiles of the remaining $n-m$ potential defectors. To see that 1-IC is equivalent to dominance note that 1-IC requires π_i to be a best response to *all* profiles with $\theta_i = \pi_i$. The definition includes the case of a profile being $(n+1)$-IC for completeness: all profiles are $(n+1)$-IC because the condition is vacuous.

Definition 2 (Critical Mass)**.** *The* critical mass *of a profile π in game u, $\kappa(\pi, u)$, is the smallest integer $m \in \{1, \ldots, n+1\}$ for which π is m-IC. For $m = 1, \ldots, n$,* the set of equilibria of critical mass at most *m, $\mathcal{C}_m(u) \equiv \{\pi \mid \pi$ is m-IC$\}$. Also, $\mathcal{C}_0(u) \equiv \emptyset$ and $\mathcal{C}_{n+1}(u) \equiv \Theta(u)$.*

The trivial solution concepts $\mathcal{C}_0$ and $\mathcal{C}_{n+1}$ are included for completeness. When the game under consideration u is clear from the context, we write $\kappa(\pi)$ for brevity.

It is easy to see that: π is a DSE iff π is 1-IC, that π is a NE iff π is n-IC, and that every profile θ is $(n+1)$-IC. It is also easy to see that π's incentive compatibility is monotonically increasing in m, i.e., for $m = 1, \dots, n$, π is m-IC implies that π is $(m+1)$-IC. Thus, $\mathcal{C}_m(u) \subseteq \mathcal{C}_{m+1}(u)$ and $\mathcal{C}_m(u) = \{\pi \mid \kappa(\pi) \leq m\}$.

The observations just made imply the nested progression of critical mass equilibria discussed in the introduction, i.e., for any game u,

$$\{\text{Dominant strategy eq}(u)\} = \mathcal{C}_1(u) \subseteq \mathcal{C}_2(u) \subseteq \cdots \subseteq \mathcal{C}_n(u) = \{\text{Nash eq}(u)\}.$$

Observation 1. *The uniformity property in the definition of incentive compatibility implies that for any m-IC profile π, at any profile θ with $m-1$ π-players, π is a uniform best response for each of the remaining $n-(m-1)$ players. More specifically, if any group G of $m-1$ players play their π strategies, then π_j is a dominant strategy for every group outsider j in the game played by members of G^c under π. This means that for any profile θ in which an entire group of m players play their π strategies it is best response for all the players of the game (including the group members themselves) to play their π strategies. This leads to the alternative description of m incentive compatibility below.*

Definition 3 (Chain Reaction)**.** *For $1 \leq m \leq n+1$, m* players incentivize a π chain reaction, *if every player's π_i strategy is a best response to any profile $\theta \in \Theta(u)$ that has at least m π-players.*

Following the observation above, we conclude that incentivizing chain reaction and (uniform) incentive compatibility are equivalent notions.

Proposition 1. *For any game u, profile π, and integer $1 \leq m \leq n+1$, π is m-IC iff m players incentivize a π chain reaction.*

The proposition above means that an alternative definition of $\kappa(\pi)$ is *the minimum number of players needed to initiate a π chain reaction.* In the terminology of Hamming balls, $\kappa(\pi)$ is the minimal integer m such that every player's π_i strategy is best response to any profile $\theta \in A_m(\pi)$.

It is also useful to consider the index ρ that is dual to the critical mass index κ:

Definition 4 (Resilience)**.** *The* index of resilience *(against defections) ρ assigns to every profile π the value $\rho(\pi) = n - \kappa(\pi)$.*

Remark 1. *The dual relationship of ρ and κ follows from the fact that "the profiles θ with d or fewer π-defectors" are exactly "the profiles θ with $n-d$ or more π-players." This duality allows us to translate statements about π-defectors to statements about π-players. Specifically, from the definition of κ, it follows that:*

(1) $\rho(\pi)$ represents the largest integer d s.t. at any profile θ with d or fewer π-defectors, π_i is a best response for every π-player; and

(2) from Proposition 1, $\rho(\pi)$ represents the largest integer d s.t. at any profile θ with d or fewer π-defectors, π_i is a best response for every player.

Remark 2. ***Relationship to Eliaz (2002)****. From Remark 1 above, it follows that an equilibrium π is d-fault-tolerant in the sense of Eliaz (2002) iff $d \leq \rho(\pi)$. But notice also that Eliaz's condition of fault-tolerance is significantly stronger: Proposition 1 shows that if the defection of the d faulty players is "tolerated at π," in the sense of Eliaz (2002), then each of these d faulty players is "disciplined at π," i.e., at any profile θ with $d \leq \rho(\pi)$ π-defectors, the defection from π_j can only lead to a loss to each defector j.*

2.2. **Critical mass in stag-hunt games.** The simplified stag-hunt games defined below offer insights into the critical-mass notion, which is useful for the proof of the main theorem. An n-player stag-hunt game is defined for every integer $t = 1, \ldots, n$.

Definition 5 (Stag Hunt)**.** An n-player stag-hunt game with threshold t, s^t: *Each player has two strategies: to participate in a hunt,* H*, or to laze,* L*. The (safe) payoff of an* L *chooser is* 0*, regardless of the opponents' choices; however the (risky) payoff of an* H *chooser is* 1 *at any profile with at least* t H*-choosers, but* -1 *at profiles in which the number of* H*-choosers is strictly smaller than* t*.*

Stag-hunt games were studied in the 1700s by the philosophers Jean-Jacques Rousseau and David Hume. Following this literature, we refer to the Nash equilibrium $\bar{H}$ in which all n players choose H as the *social contract.* In playing $\bar{H}$, every player chooses the risky action that would yield them the highest possible payoff, but only if at least $t-1$ of their opponents also choose this risky action. For this reason, stag-hunt games are sometimes thought of as games of trust. Also, for this reason the stag hunt has critical mass $\kappa(\bar{H}, s^t) = t$. For a broader view, we refer the reader to Skyrms (2001).

In the proof of the main theorem of this paper, the social contracts of stag-hunt games serve as benchmarks to all the equilibrium concepts in the vNN framework (see Section 3.2). An experimental study (Kim, Min, and Wooders, 2022) show, with statistical significance, that players' participation in the social contract decreases as the required critical mass of the equilibrium (the threshold) increases.

3. Characterization of critical mass equilibrium

In this section we present the formal theorem showing that an equilibrium concept satisfies three minimal properties (axioms) of the von Neumann and Nash (vNN) optimization framework iff it is one of the critical mass concepts $\mathcal{C}_m$ described in the previous section. These axioms are slightly more complex than previous axiomatizations of NE (Salonen, 1992; Kaneko, 1994). Moreover, since the "if" direction of the characterization is the more applicable and challenging part, choosing the smallest number of weakest possible axioms makes the theorem more significant. This does not mean that properties beyond what is described in this section are of no interest.

Several such stronger useful properties are satisfied by the $\mathcal{C}_m$'s, and some are not. But all three axioms are essential for the proof of the theorem.

The chosen axioms describe the properties of the two main solution concepts of non-cooperative game theory: Dominant strategy equilibrium and Nash equilibrium. In this sense the application of these axioms is valid on any domain of strategic games in which the use of DSE and NE is valid. Such domains may specify games with finite versus infinite number of strategies, may allow pure strategies versus mixes strategies, etc.

3.1. **Three minimal axioms from the vNN framework.**

To describe the axioms and the proof that follows, we define the *best-response justifications* for player i playing a particular strategy α_i to be the opponent profiles to which α_i is a (weak) best response; more formally,

$$J_i(\alpha_i, u_i) \equiv \{\theta_{-i} \in \Theta_{-i}(u_i) \mid u_i(\alpha_i, \theta_{-i}) \geq u_i(\beta_i, \theta_{-i}) \text{ for all } \beta_i \in \Theta_i(u_i)\}\,. \tag{3.1}$$

These sets were used earlier by Harsanyi and Selten (1988), who referred to them as the *stability sets*. We use the term *justification* to remind the reader of the thinking of a best-response strategy chooser. Notice also that a profile π of a game u is m-IC iff, for every player i, $\theta_{-i} \in J_i(\pi_i, u_i)$ for every profile θ with at least m π-players; equivalently, using the agreement ball defined in Eq. (2.2):

$$\kappa(\pi) \leq m \text{ iff } A_{m-1}(\pi_{-i}) \subseteq J_i(\pi_i, u_i). \tag{3.2}$$

3.1.1. *Best Response Monotonicity.* The first axiom, Best Response Monotonicity (BRM), states that an equilibrium concept should be monotonic in J_i. In particular, if the payoff of one player i is modified to u_i', so that their equilibrium strategy π_i is a best response to even more of their opponents' profiles, then the equilibrium remains.

Axiom 1 (BRM)**.** *Let $u, u' \in \mathcal{U}_n$ with $\Theta(u) = \Theta(u')$, and let $i \in N$ such that for all $j \neq i$, $u_j = u'_j$. If $\pi \in \mathcal{E}(u)$ and $J_i(\pi_i, u_i) \subseteq J_i(\pi_i, u'_i)$, then $\pi \in \mathcal{E}(u')$.*

Clarification: it is important to recognize that the expanded justification at π_i , i.e., $J_i(\pi_i, u_i) \subseteq J_i(\pi_i, u'_i)$ does not imply expanded justifications at other closely related strategies. For example, it does not mean that $J_i(m_i, u_i) \subseteq J_i(m_i, u'_i)$ in which m_i is any mixed strategy that assigns positive probability to π_i.

The BRM axiom above is applicable to changes in the payoffs of one player. However, it is equivalent to simultaneous changes of payoffs of more than one player. This follows from the fact that a player's justification set $J_i(\pi_i, u_i)$ is a function only of that player's payoff u_i.

Observation 2 (Simultaneous BRM)**.** *Let $\mathcal{E}$ be an equilibrium concept satisfying BRM. Let $u, u' \in \mathcal{U}_n$ with $\Theta(u) = \Theta(u')$, and let $\pi \in \mathcal{E}(u)$ be such that $J_i(\pi_i, u_i) \subseteq J_i(\pi_i, u'_i)$ for all $i \in N$. Then it also holds that $\pi \in \mathcal{E}(u')$.*

Proof. Consider intermediate games $u = u^0, u^1, u^2, \ldots, u^n = u'$, where game u^i is the same as game u^{i-1} except that player i's payoffs are updated to u'_i. These games all satisfy $\pi \in \mathcal{E}(u^i)$ by BRM, since $J_i(\pi_i, u_i^{i-1}) = J_i(\pi_i, u_i) \subseteq J_i(\pi_i, u'_i) = J_i(\pi_i, u_i^i)$. ∎

BRM has other important consequences. For instance, it implies a scale invariance property as well, where we scale all payoffs by a positive constant.

Observation 3 (Scale Invariance)**.** *Consider any constant $c > 0$ and games u, u' with $\Theta(u) = \Theta(u')$ and $u_i(\theta) = cu'_i(\theta)$ for all $i \in N$, $\theta \in \Theta(u)$. Then $\mathcal{E}(u) = \mathcal{E}(u')$.*

Proof. This follows from Simultaneous BRM because rescaling by a positive constant does not change the justification sets. ∎

3.1.2. *Sure Thing Principle.* Following common terminology, the axiom below is called the *Sure Thing Principle.* It states that if the same profile π is an equilibrium in two different games u, u', then it is also an equilibrium in a game in which nature randomizes which of the two games is played. That is to say, if nature flips a coin with bias λ to determine whether game u or u' is played and players have to choose a strategy to use in either game, then it is an equilibrium in this random game for them to choose any equilibrium common to the two component games. However, as illustrated below, it may be the case that new equilibria are introduced.

Axiom 2 (Sure Thing Principle). *For any $u, u' \in \mathcal{U}_n$ with $\Theta(u) = \Theta(u')$ and any $0 \leq \lambda \leq 1$, it must be the case that $\mathcal{E}(u) \cap \mathcal{E}(u') \subseteq \mathcal{E}(u'')$, where $u'' : \Theta(u) \to \mathbb{R}^n$ is defined by $u''(\theta) \equiv \lambda u(\theta) + (1 - \lambda)u'(\theta)$.*

The Sure Thing Principle is also motivated by best response, in the following sense. Both NE and DSE can be defined as mutual best responses, for different notions of optimality. The Sure Thing Principle in Decision Theory states that if a decision is optimal in the case of an event E as well as its complement $\neg E$, then it is optimal in any event (Savage, 1954). Applying this principle separately to the response optimality of each player yields Axiom 2.

Example 3. *New equilibria emerge under uncertainty, but old uniform ones are not eliminated. Consider the following three games:*

Sunny-day game			***Rainy-day game***			***50/50-day game***		
	walk	*eat*		*walk*	*eat*		*walk*	*eat*
walk	*4, 0*	*0, 0*	*walk*	*0, 4*	*0, 0*	*walk*	*2, 2*	*0, 0*
eat	*0, 0*	***1, 1***	*eat*	*0, 0*	***1, 1***	*eat*	*0, 0*	***1, 1***

These two-player games have payoffs 0 if players do not coordinate on the same action. Player 1 likes to walk in the sun while player 2 likes to walk in the rain. If the unknown weather is equally likely to be rain or sun, then the expected payoffs are those of the 50/50 game. The Sure Thing Principle implies that, if $\pi = (eat, eat)$ is an equilibrium in the Sunny and Rainy games, then π must also be an equilibrium in the 50/50 game.

In the 50/50 game, one can imagine (walk, walk) emerging as a new equilibrium. The Sure Thing Principle, however, says that π is not eliminated. This is justified by best response: if each player is motivated by best responses, and eat is an optimal response in both games, then it must be an optimal response in the 50/50 game. Put another way, "making deals" is not required for a best-responder. There may be reasonable equilibrium concepts that exclude (eat, eat) in the 50/50 game, but such concepts must use reasoning outside of a best response framework.

3.1.3. *Anonymity.* Under a variety of names, this axiom has been used in cooperative game theory and in social choice. It states that equilibria should be invariant to the names of players and to the labels and replication of strategies.

Axiom 3 (Anonymity)**.** *Let $u \in \mathcal{U}_n$.*
(1) Player anonymity: *For permutation $\rho : N \to N$, let $\rho(\theta) \equiv \big(\theta_{\rho(1)}, \ldots, \theta_{\rho(n)}\big)$, and let $u' \in \mathcal{U}_n$ be the game defined as follows:*

$$u'_{\rho(i)}(\theta) \equiv u_i(\rho(\theta)) \text{ for all } i \in N, \theta \in \Theta(u) \text{ and } \Theta_{\rho(i)}(u') \equiv \Theta_i(u).$$

Then $\mathcal{E}(u') = \{\pi \in \Theta(u') \mid \rho(\pi) \in \mathcal{E}(u)\}$.
(2) Strategy anonymity: *Let $u' \in \mathcal{U}_n$ and let $\tau_i : \Theta_i(u') \to \Theta_i(u)$ for $i \in N$ be surjective (onto) functions such that:*

$$u'(\theta) = u\big(\tau_1(\theta_1), \ldots, \tau_n(\theta_n)\big) \text{ for all } \theta \in \Theta(u').$$

Then, $\mathcal{E}(u') = \big\{\theta \in \Theta(u') \mid \big(\tau_1(\theta_1), \ldots, \tau_n(\theta_n)\big) \in \mathcal{E}(u)\big\}$.

The surjective requirement in (2) guarantees that the game u' is formed by renaming and possibly replicating strategies. Equivalently, (2) can be written as two separate properties for any given player i: (2a) renaming its strategies with a bijection between $\Theta_i(u')$ and $\Theta_i(u)$, and (2b) duplicating strategies defined as follows:

Observation 4 (Strategy Duplication)**.** *Let $\mathcal{E}$ be an equilibrium concept satisfying Anonymity. Let $u, u' \in \mathcal{U}_n$ and $\tau_i : \Theta_i(u') \to \Theta_i(u)$ for $i \in N$ be such that $\Theta(u) \subseteq \Theta(u')$ and:*

$$\tau_i(\theta_i) = \theta_i \textit{ for all } \theta_i \in \Theta_i(u); \textit{ and}$$
$$u'(\theta) = u\left(\tau_1(\theta_1), \ldots, \tau_n(\theta_n)\right) \textit{ for all } \theta \in \Theta(u').$$

Then, $\mathcal{E}(u') = \{\theta \in \Theta(u') \mid \big(\tau_1(\theta_1), \ldots, \tau_n(\theta_n)\big) \in \mathcal{E}(u)\}$.

Proof. This follows trivially as a special case of Strategy Anonymity, because the above τ_i are clearly surjective. ∎

Moreover, Anonymity also implies that one can remove redundant strategies, as shown in Observation 5 below.

Observation 5 (Redundant Strategy Elimination)**.** *Let $\mathcal{E}$ be an equilibrium concept satisfying Anonymity. Let $u, u' \in \mathcal{U}_n$ and let $\tau_i : \Theta_i(u) \to \Theta_i(u')$ for $i \in N$ be such that $\Theta(u') \subseteq \Theta(u)$ and:*

$$\tau_i(\theta_i) = \theta_i \textit{ for all } \theta_i \in \Theta_i(u'); \textit{ and}$$
$$u(\theta) = u'\big(\tau_1(\theta_1), \ldots, \tau_n(\theta_n)\big) \textit{ for all } \theta \in \Theta(u).$$

Then, $\mathcal{E}(u') = \mathcal{E}(u) \cap \Theta(u')$.

Proof. One can view the removal of redundant strategies as a duplication going from u' to u. In particular, apply the previous Observation 4, but swap u and u'. The observation states that $\mathcal{E}(u) = \{\theta \in \Theta(u) \mid (\tau_1(\theta_1), \ldots, \tau_n(\theta_n)) \in \mathcal{E}(u')\}$. Hence,

$$\mathcal{E}(u) \cap \Theta(u') = \{\theta \in \Theta(u') \mid \theta \in \mathcal{E}(u')\} = \mathcal{E}(u'),$$

because $\tau_i(\theta_i) = \theta_i$ for all $\theta_i \in \Theta_i(u')$. ∎

3.2. **Main theorem, statement, and proof.** We now proceed with the main theorem of this paper. Our analysis includes the trivial equilibrium concepts $\mathcal{C}_0 \equiv \emptyset$ and $\mathcal{C}_{n+1} \equiv \Theta$ because their inclusion paints a complete picture, but of course they could be excluded by adding an additional axiom which states that the solution concept cannot be trivial.

Theorem 1. *An equilibrium concept $\mathcal{E}$ satisfies Best-Response Monotonicity (BRM), the Sure Thing Principle, and Anonymity iff $\mathcal{E} = \mathcal{C}_m$ for some $m \in \{0, 1, 2, \ldots, n+1\}$.*

To prove the theorem, we first argue that each $\mathcal{C}_m$ satisfies the three axioms and then, more interestingly, argue that they are the unique equilibrium concepts that satisfy these axioms.

Lemma 1. *The concepts $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_{n+1}$ are distinct and satisfy Axioms 1-3.*

Proof. As discussed in the section on stag-hunt games, (Section 2.2), the fact that $\bar{H} \in \mathcal{C}_t(s^t) \setminus \mathcal{C}_{t-1}(s^t)$ for the stag-hunt games with $t = 1, 2, \ldots, n+1$ shows that $\mathcal{C}_0, \ldots, \mathcal{C}_{n+1}$ are distinct. Next we argue that $\mathcal{C}_m$ satisfies Axioms 1-3 for any m.

Each $\mathcal{C}_m$ clearly satisfies BRM because $\mathcal{C}_m$ is equivalent to the best-response justification sets containing an agreement ball (see Eq. 3.2). This certainly remains true if the justification sets are enlarged. To see that $\mathcal{C}_m$ satisfies the Sure Thing Principle, assume that $\pi \in \mathcal{C}_m(u) \cap \mathcal{C}_m(u')$, i.e., each π_i is a best-response in both u and u' against any profile θ that involves at least $m-1$ opponents playing π. But the notion of best-responses satisfies the Sure Thing Principle, implying that for any $0 \le \lambda \le 1$, each π_i is a best response in $\lambda u + (1-\lambda)u'$ against any such profile, and thus $\pi \in \mathcal{C}_m(\lambda u + (1-\lambda)u')$. Finally, the definition of $\mathcal{C}_m$ is clearly symmetric to player order. Moreover, duplicating or renaming strategies does not affect best responses; hence, $\mathcal{C}_m$ satisfies Anonymity. ∎

The rest of this section is devoted to proving the other direction of the main theorem: that any equilibrium concept $\mathcal{E}$ that satisfies Axioms 1-3 must be one of $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_{n+1}$. In particular, we prove that $\mathcal{E} = \mathcal{C}_{M(\mathcal{E})}$ for the *most fragile equilibrium* $M(\mathcal{E})$ defined by

$$M(\mathcal{E}) \equiv \max\left(\{\kappa(\pi, u) \mid u \in \mathcal{U}_n,\ \pi \in \mathcal{E}(u)\} \cup \{0\}\right). \tag{3.3}$$

If $\mathcal{E}(u)$ is empty for all u, then $M(\mathcal{E}) = 0$.

The proof makes use of two propositions that relate those equilibrium concepts satisfying the axioms to the social-contract equilibria $\bar{H}$ (everybody hunts) in the stag-hunt games with threshold t, s^t. We will henceforth assume that $\{H, L\} \subseteq \overline{\Theta}$ are identified with two possible strategies contained in the superset of all possible strategies. This is possible since we have assumed that there are at least $|\overline{\Theta}| \geq 3$ possible strategies.

Recall that for the game s^t, $\kappa(\bar{H}, s^t) = t$, as argued in Section 2.2. The easier of the two propositions states that if such an $\bar{H}$ is an $\mathcal{E}$-equilibrium of s^t, then all the profiles π of games u in which $\kappa(\pi) \leq t$ must be $\mathcal{E}$-equilibria in their games.

Proposition 2. *Let $\mathcal{E}$ be an equilibrium concept satisfying BRM and Anonymity. Then for any $1 \leq t \leq n+1$, if $\bar{H} \in \mathcal{E}(s^t)$, then $\mathcal{E}(u) \supseteq \mathcal{C}_t(u)$ for all $u \in \mathcal{U}_n$.*

Proof. Fix any $1 \leq t \leq n+1$ and suppose $\bar{H} \in \mathcal{E}(s^t)$. For any $u \in \mathcal{U}_n$ and $\pi \in \mathcal{C}_t(u)$, we must show that $\pi \in \mathcal{E}(u)$. Fix such a $\pi \in \mathcal{C}_t(u)$. WLOG, we can assume that $\Theta_i(u)$ has at least two strategies for each player i. This follows from Anonymity, because for any player who has only π_i as a strategy, we can duplicate π_i to create a second equivalent strategy without affecting π's membership in $\mathcal{E}(u)$ or $\mathcal{C}_t(u)$, by Observation 4.

Next, we consider a hybrid game $u' \in \mathcal{U}_n$ between the game s^t and u, with u''s profiles $\Theta(u') = \Theta(u)$ but whose payoff function u' "mimics" s^t:

$$u'_i(\theta) = \begin{cases} 1 & \text{if } \theta_i = \pi_i \text{ and } a(\theta,\pi) \geq t \\ 0 & \text{if } \theta_i \neq \pi_i \\ -1 & \text{if } \theta_i = \pi_i \text{ and } a(\theta,\pi) < t. \end{cases}$$

The games u' and s^t are equivalent up to renaming π_i to H, and every other strategy to L. Because we have ensured that each player has at least two strategies, at least one strategy corresponds to L. Thus, by Anonymity, since $\bar{H} \in \mathcal{E}(s^t)$, it follows that $\pi \in \mathcal{E}(u')$.

Next, we claim:

$$J_i(\pi_i, u'_i) = A_{t-1}(\pi_{-i}) \subseteq J_i(\pi_i, u_i).$$

The first equality holds by definition of u'_i and A_{t-1} (see Eq. 2.2). The second follows from the definition of $\mathcal{C}_t$ and the fact that $\pi \in \mathcal{C}_t(u)$. Since $\pi \in \mathcal{E}(u')$, simultaneous BRM (see Observation 2) implies $\pi \in \mathcal{E}(u)$. ∎

A more difficult direction states involves stating a strong converse to the proposition above: if $\pi \in \mathcal{E}(u)$ for some u, then $\bar{H}$ is an equilibrium for the game $s^{\kappa(\pi)}$. An equivalent statement is described in the next proposition.

Proposition 3. *Let $\mathcal{E}$ be an equilibrium concept satisfying Axioms 1-3. Let $m = M(\mathcal{E})$ as defined in Eq. (3.3). If $m \geq 1$, then $\bar{H} \in \mathcal{E}(s^m)$.*

It is not difficult to see that these two propositions imply the main theorem. Before we present the proof of Proposition 3, we use it and Proposition 2 to prove the main theorem.

Proof of Theorem 1. First, if $m = 0$ so that $\mathcal{E}(u) = \emptyset$ for all games u, then $\mathcal{E} = \mathcal{C}_0$, and we are done. Otherwise, by Prop. 3, $\bar{H} \in \mathcal{E}(s^m)$ and thus Prop. 2 implies that

$\mathcal{E}(u) \supseteq \mathcal{C}_m(u)$ for all u. On the other hand, $\mathcal{E}(u) \subseteq \mathcal{C}_m(u)$ by the definition of M; otherwise, there would be a more fragile equilibrium. ∎

The remainder of this section is devoted to the proof of Proposition 3. To this end, it will be helpful to define, for each $1 \leq t \leq n+1$, the game h^t, which is somewhat simpler than s^t. As stepping stones for the proof, we use additional games f^t and g^t. The games are all defined on the same sets of strategies, specifically $\Theta(f^t) \equiv \Theta(g^t) \equiv \{H, L\}^n$ with, for all $i \in N$:

$$h_i^t(\theta) \equiv \begin{cases} 1 & \text{if } \theta_i = L \text{ and } a(\theta, \bar{H}) \leq t-2 \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } \bar{H} \equiv (H, H, \ldots, H); \tag{3.4}$$

$$g_i^t(\theta) \equiv \begin{cases} 1 & \text{if } \theta_i = L \text{ and } a(\theta, \bar{H}) = t-2; \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

$$f_i^t(\theta) \equiv \begin{cases} 1 & \text{if } i = 1 \text{ and } \theta = \delta^t \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } \delta^t \equiv (\underbrace{L, L, \ldots, L}_{n-t+2}, \underbrace{H, H, \ldots, H}_{t-2}) \tag{3.6}$$

Note that for $t = 1$, the three games are all defined to have identically 0 payoffs. Also observe that the hunting equilibrium in h^t and s^t is similar in the sense of Observation 6 below.

Observation 6. *Let $\mathcal{E}$ be an equilibrium concept satisfying BRM, and let $1 \leq t \leq n+1$. Then $\bar{H} \in \mathcal{E}(h^t)$ if and only if $\bar{H} \in \mathcal{E}(s^t)$.*

Proof. The two games have identical strategies: $\Theta(h^t) = \Theta(s^t) = \{H, L\}^n$. We claim they also have identical justification sets:

$$J_i(H, h_i^t) = J_i(H, s_i^t) = A_{t-1}(\bar{H}_{i-1}).$$

To see this for h^t, notice that as long as at least $t-1$ opponents play H, a player's payoff will be 0 regardless; hence, H is a best response. On the other hand, if fewer than $t-2$ opponents play H, then H is not a best response. A similar argument applies for s^t. Thus, by BRM, $\bar{H} \in \mathcal{E}(s^t)$ if and only if $\bar{H} \in \mathcal{E}(h^t)$. ∎

The proof of Proposition 3 is structured as follows. Observation 6 means that to prove Proposition 3, we need to show only that $\bar{H} \in \mathcal{E}(h^m)$ for $m = M(\mathcal{E})$. We do this by first showing that $\bar{H} \in \mathcal{E}(f^t)$ for $t \leq m$ and how this implies that $\bar{H} \in \mathcal{E}(g^t)$ for all $t \leq m$ and ultimately that $\bar{H} \in \mathcal{E}(h^m)$.

Lemma 2. *Let $\mathcal{E}$ be an equilibrium concept satisfying BRM and Anonymity, such that $M(\mathcal{E}) \geq 1$. Then $\bar{H} \in \mathcal{E}(f^t)$ for all $t \in \{1, 2, \ldots, m\}$.*

Note that the proof of this lemma is the only place where we use the assumption that there at least $|\overline{\Theta}| \geq 3$ strategies on which games can be defined.

Proof. Let $m = M(\mathcal{E}) \in \{1, 2, \ldots, n+1\}$ and let u and $\pi \in \mathcal{E}(u)$ be such that $m = \kappa(\pi, u)$, which must exist by definition of M. By definition of κ, there must be some player i and some defection profile γ such that $u_i(\gamma) > u_i(\pi_i; \gamma_{-i})$, with $a(\gamma_{-i}, \pi_{-i}) = m-2$; otherwise, $\kappa(\pi, u) < m$ (i.e., π would not be so fragile).

WLOG, by Anonymity, we can permute players and rename strategies so that $i = 1$ and

$$\pi = \bar{H} = (\underbrace{H, H, \ldots, H}_{n});$$

$$\gamma = \delta^m = (\underbrace{L, L, \ldots, L}_{n-m+2}, \underbrace{H, H, \ldots, H}_{m-2}); \text{ and}$$

$$u_1(\gamma) > u_1(H; \gamma_{-1}).$$

WLOG, we can further assume that all players i have at least these two strategies, $\{H, L\} \subseteq \Theta_i$, among others. Clearly all players have a strategy H, and for each player who does not already have strategy $L \in \Theta_i$, one could duplicate strategy H to form a new strategy named L which would keep $\bar{H} \in \mathcal{E}(u)$ by Anonymity.

Now, consider the hybrid game v between u and f^m which has the profiles $\Theta(v) = \Theta(u)$ of u but has all payoffs 0 except $v_1(\delta^m) = 1$. It follows from BRM that $\bar{H} \in \mathcal{E}(v)$ because the only situation in which H is not a best response in v is for player 1 at profile δ^m, but in that case H was also not a best response in u. Hence, the justification sets $J_i(H, v_i) \supseteq J_i(H, u_i)$ for each player, and Simultaneous BRM (see Observation 2) implies that $\bar{H} \in \mathcal{E}(v)$ because $\bar{H} \in \mathcal{E}(u)$.

Next, we observe that each player's strategies in v are equivalent to either H or L, and thus the game can be reduced down to these two strategies per player, which is exactly the game f^m. This holds even in the case of infinitely many strategies. This is simply a matter of eliminating redundant strategies, which preserves equilibrium $\bar{H}$ by Observation 5.

We have thus argued that $\bar{H} \in \mathcal{E}(f^m)$ and it remains to show that $\bar{H} \in \mathcal{E}(f^t)$ for all $1 \leq t \leq m$. To see this, consider any such t. We will create another hybrid game w between f^t and f^m. Starting with the game f^m, consider the game h formed by taking f^m and for the $m - t$ players $n - m + 2 \leq i \leq n - t + 2$: first, rename strategy L to L' (which is possible since we have assumed there are at least $|\overline{\Theta}| \geq 3$ strategies) and then create a duplicate strategy of H named L. By Anonymity, $\bar{H} \in \mathcal{E}(w)$, and by design $w(\delta^t) = 1$. Now, we will use an approach similar to the one above to transform the game w to f^t while preserving $\bar{H} \in \mathcal{E}(f^t)$. Specifically, consider changing the payoffs in w so that they are all 0 except $w(\delta^t) = 1$. It again follows from BRM that $\bar{H}$ remains an equilibrium, according to $\mathcal{E}$, because the only situation in which H is not a best response in this new game is for player 1 at profile δ^t, but in that case H was also not a best response in w. And again the extra strategy L' is equivalent to L in the new game, so it can be eliminated to form exactly the game f^t. ∎

Using the Sure Thing Principle, and averaging f^t over permutations of players, Lemma 3 shows the equivalence of $\bar{H} \in \mathcal{E}(f^t)$ and $\bar{H} \in \mathcal{E}(g^t)$.

Lemma 3. *Let $\mathcal{E}$ be an equilibrium concept satisfying Axioms 1-3. For each $1 \leq t \leq n+1$, $\bar{H} \in \mathcal{E}(f^t)$ if and only if $\bar{H} \in \mathcal{E}(g^t)$.*

Proof. Fix $1 \leq t \leq n+1$, and for shorthand define $u = f^t$. First, by BRM, it is easy to see that if $\bar{H} \in \mathcal{E}(g^t)$, then $\bar{H} \in \mathcal{E}(u)$, because $\bar{H}$ is more justified in u than in g^t. In particular, H is always a best response in u except for player 1 at δ^t, but in that case H is also not a best response in g^t.

Next, suppose that $\bar{H} \in \mathcal{E}(u)$. It remains to show $\bar{H} \in \mathcal{E}(g^t)$. We consider the game $\bar{u}$ defined by averaging the payoffs of u over all permutations of players. Formally, let $\Pi(N)$ denote the set of permutations of players. For $\rho \in \Pi(N)$, define the ρ-permuted game $u^\rho : \{H, L\}^n \to \{0, 1\}$ by

$$u^\rho_{\rho(i)}(\theta) \equiv u_i(\theta_{\rho(1)}, \theta_{\rho(2)}, \ldots, \theta_{\rho(n)}) \text{ for all } \theta \in \Theta(u^\rho) = \{H, L\}^n.$$

By Player Anonymity, $\bar{H} \in \mathcal{E}(u^\rho)$ for each $\rho \in \Pi(N)$. Let $\bar{u}$ be the average payoff in these games which are all on strategy profiles $\{H, L\}^n$, more formally:

$$\bar{u}_i(\theta) = \frac{1}{n!} \sum_{\rho \in \Pi(N)} u^\rho_i(\theta) \text{ for all } i \in N, \theta \in \{H, L\}^n.$$

By the Sure Thing Principle, $\bar{H} \in \mathcal{E}(\bar{u})$. It is easy to see by symmetry that, for some constant $c > 0$:

$$\bar{u}_i(\theta) = \begin{cases} c & \text{if } \theta_i = L \text{ and } a(\theta, \bar{H}) = t - 2, \\ 0 & \text{otherwise.} \end{cases}$$

This is because a constant fraction of the permutations will result in player i being mapped to player 1 and the H-players being the last $t-2$ players. Since $\bar{u}$ is simply a factor-c rescaling of g^t, Scale Invariance (Observation 3 using BRM) implies $\bar{H} \in \mathcal{E}(g^t)$ as well. ∎

With these lemmas, we can now prove Proposition 3. The best-response structure of g^t and h^t differ; in particular, H is *not* a best response in the game when any $t-2$ *or fewer* players play H while, in g^t, H is only suboptimal when exactly $t-2$ players play H. We use the Sure Thing Principle to average the payoffs over $g^1, \ldots, g^t$ to prove Proposition 3.

Proof of Proposition 3. Let $m \in M(\mathcal{E}) \in \{1, 2, \ldots, n+1\}$. By Lemma 2, $\bar{H} \in \mathcal{E}(f^t)$ for all $t \in \{1, 2, \ldots, m\}$. By Lemma 3, we also have $\bar{H} \in \mathcal{E}(g^t)$ for $t \in \{1, \ldots, m\}$. Consider the average of these games:

$$\bar{g} \equiv \frac{1}{m}(g^1 + g^2 + \ldots + g^m).$$

By the Sure Thing Principle, $\bar{H} \in \mathcal{E}(\bar{g})$. Observe that $\bar{g} = \frac{1}{m}h^m$ is equivalent to the game h^m with its payoffs scaled down by a factor m. By Scale Invariance (Observation 3), it follows that $\bar{H} \in \mathcal{E}(h^m)$. Finally, by Observation 6, this in turn implies that $\bar{H} \in \mathcal{E}(s^m)$. ∎

3.3. **Violations of the vNN axioms:** Many equilibrium concepts violate the Anonymity axiom. A simple example is an equilibrium concept that is *dictatorial*, e.g., π is an equilibrium iff $u_1(\pi) \geq u_1(\theta)$ for every profile θ. It is important to note the difference between a dictatorial equilibrium concept and a game with a dictator. The critical mass concept introduced in this paper, which is a non-dictatorial solution concept, is still applicable to games in a community that is controlled by a dictator, as in our Centralized Chip Production game, Example 2 in the introduction.

To illustrate a violation of the best response monotonicity axiom, consider the notion of 0.1-Nash equilibrium (0.1-NE), and the two 2×2 pure strategy games $u \equiv \begin{array}{|c|c|}\hline 0.95,0 & 1,0 \\ \hline 1.00,0 & 0,0 \\ \hline \end{array}$ and $u' \equiv \begin{array}{|c|c|}\hline 0.85,0 & 1,0 \\ \hline 1.00,0 & 0,0 \\ \hline \end{array}$. The justification set of the top strategy of the row chooser consist of right side strategy of the column chooser in both games, yet the top left profile (T,L) is a 0.1-NE in the first game but not in the second. Notice however that games may exhibit a difference in the justification sets if they include mixed strategies.

The best response monotonicity axioms imply that we restrict ourselves to ordinal considerations when deciding whether a profile is an equilibrium. As an example, the cardinal considerations needed for the identification of ε-NE (for any fixed $\varepsilon > 0$) profiles rule out ε-NE as a possible equilibrium concept within the standard von Neumann-Nash framework.

The following example illustrates a limitation of the BRM axiom, like all ordinal concepts including NE and DS.

Example 4. *Let $g = 10^{100}$ be the number googol. Consider the following two games:*

$$\begin{array}{|c|c|}\hline 0,0 & 0,-\varepsilon \\ \hline g,0 & g,g \\ \hline \end{array} \qquad \begin{array}{|c|c|}\hline 0,0 & 0,-g \\ \hline \varepsilon,0 & \varepsilon,\varepsilon \\ \hline \end{array}$$

The two games have identical best-response structure and identical justification sets. However, the NE profile (B,R) is more compelling in the LHS game because the column chooser stands to lose at most ε or gain g by playing R, whereas R in the RHS game may cost the column chooser more than g while it can lead to a gain of at most ε. Thus, "playing it safe" by playing L is more tempting in the RHS game. BRM, however, implies that if (B,R) is an equilibrium in LHS then it also must be an equilibrium in the RHS game. Such ordinal reasoning, which disregards the magnitudes of differences, is a weakness of BRM axiom. However, it is also a limitation of NE, which has nonetheless proven to be a useful solution concept.

The same issue arises with dominant strategies, as can be seen in the following two prisoner's dilemma games, where the DSE (B,R) is again arguably more compelling the LHS game:

3,3	0,4
4,0	1,1

3,3	$-g-\epsilon$,$3+\epsilon$
$3+\epsilon,-g-\epsilon$	$-g,-g$

The well-known "trembling hand" perfect equilibrium (PE) of Selten is an equilibrium concept that violates the Sure-Thing Principle, as shown next.

Consider first the two-player 2×3 game Γ in which the payoffs of player 2 are identically zero and the payoffs of player 1 are described by the following table:

	L	M	R
T	.5	1	.5
B	1	1	1
	ε	$1-\varepsilon-\delta$	δ

It is clear that, under any trembles of player 2 (i.e., positive ε or δ in the bottom row), T is a dominated strategy. Thus, *the profile* (T, M) *is not a perfect equilibrium.*

Now consider two component games Γ_L and Γ_R described by the tables below, in which player 2 payoffs are zero again.

	L	M	R
T	1	1	0
B	0	1	2
	4ε	$1-5\varepsilon$	ε

	L	M	R
T	0	1	1
B	2	1	0
	ε	$1-5\varepsilon$	4ε

The trembles described in the bottom rows of these component games show that *the profile* (T, M) *is a perfect equilibrium in both component games.* This exhibits a violation of the Sure-Thing Principle since the game $\Gamma = .5\Gamma_L + .5\Gamma_R$.

The violation above suggests a type of deficiency of perfect equilibrium as a solution concept. Determining whether a profile is a perfect equilibrium depends on information outside the data of the game under consideration (i.e., the payoffs of the players). More specifically, it depends on imaginary trembles that an analyst may assign to component games to fit situations in which the game is played.

4. Applications and further illustrations

This section elaborates on the notions of critical mass and their applicability by examining their performance on both theoretical and actually observed games. We start with broadly applicable families of games in which the critical mass concepts have immediate natural interpretations.

The first family illustrates the fragility of mixed strategies, used in many applications. As illustrated in the game below, whenever mixed strategies are meaningful, the mixed strategy equilibrium is fragile.

4.1. **A Game with mixed strategies.**

Consider three competitive sellers $s = 1, 2, 3$. Simultaneously each has to choose to participate in one of two possible markets, M_x, with $x = 1$ or 2. Whatever market is chosen by s, s's payoff is $u_s = 1/N_x$, where N_x is the number of the sellers (including s) who participate in the market M_x. Thus, if s is the only seller in a market then her payoff is 1, if she is in a market with one other seller then her payoff is 1/2, and if she is in a market with the two other sellers then her payoff is 1/3. We consider the symmetric mixed strategy Nash equilibrium of this market choice game described by the profile π in which every seller $s = 1, 2, ...3$ randomizes, $\pi_s(M_1) = \pi_s(M_2) = 0.5$.

It is easy to check that if seller 1 change her mixed strategy to be $\pi'_1(M_1) = 0.5 + \epsilon$ for any positive ϵ and seller 2 stays with $\pi'_2(M_1) = 0.5$, then the best response of seller 3 is to choose M_2 with certainty, i.e. $\pi'_3(M_2) = 1$. Thus, the mixed strategy is (very) fragile since a change of probabilities (no matter how small) by one player incentivizes another seller to defect.

The next family of games provide natural interpretations for the critical mass index, κ.

4.2. **Participation games.** An n-person participation game is described by two parameter profiles $(s,t) = (s_i,t_i)_{i=1,...,n}$. Every player i has two strategies: a *risky* strategy P that describes *participation* in a certain group activity, and a *safe* strategy A that means *avoid participation.* In any profile θ with $\theta_i = A$ player i's payoff is $u_i(\theta) = s_i$, where s_i is a real number referred to as i's *safe payoff.* On the other hand, in any profile θ with $\theta_i = P$, player i's payoff depends on the number of participants in θ (including player i), defined by $\#P(\theta) \equiv |\{j \in N : \theta_j = P\}|$: For the integer $t_i = 1, 2, ..., n$, referred to as player i's *participation threshold*: if $\#P(\theta) \geq t_i$, then $u_i(\theta) > s_i$; but if $\#P(\theta) < t_i$, then $u_i(\theta) < s_i$.

The social contract is the profile of full participation $\overline{P}$. For $t \equiv \max_i(t_i)$, it is easy to see that $\kappa(\overline{P}) = t$; in other words, the critical mass needed for full participation is the participation threshold of the most reluctant participant(s).

The stag hunt games s^t defined in Section 2.2 are participation games in which P denotes participation in the joint stag hunt and L denotes nonparticipation. The minimal number of hunters needed for a successful hunt of the particular stag being hunted, t, is the common threshold of all the hunters, i.e., $t_i = t$ for all i. The safe payoff is zero. For $t = 1, 2, ..., n$, the n social contracts $(\overline{P}^t)_{t=1,...n}$ ($\overline{P}$'s of the games s^t), represent n distinct profiles in the n corresponding equilibrium concepts $\mathcal{C}_t$. The proof of the main theorem is based on this representation.

The Rebellion game in the introduction (see Example 1) is a participation game in which the risky strategy is to rebel and the safe strategy is to acquiesce. The safe payoff is zero, and $t_i = 2$ for all i. Thus, full participation has critical mass 2. Similar simple games are illustrated next.

Party RSVP games. Every player may choose P as a positive response to a party invitation, or choose H as a negative response. Every player strictly prefers P to H in any profile in which she is not the only P chooser, but strictly prefers H to P if she is the only P chooser. These are participation games in which P is the risky strategy, all $t_i = 2$, and s_i denotes the payoff of a player who chooses to stay home. It is easy to see that all invitees attending the party has critical mass $\kappa(\overline{P}) = 2$: thus, $\overline{P}$ is a nearly dominant strategy equilibrium. Notice also that the other pure strategy equilibrium $\overline{H}$, in which everybody chooses H, is a fragile Nash equilibrium: One defector from $\overline{H}$ to P incentivize the others to defect from $\overline{H}$ to P.

A **Bonus game** exemplifies a game in which full participation is a fragile equilibrium. In this game, each member of an n-person production team may exert extra effort, E, or play the lazy strategy, L. If all play E, they each receive a valuable bonus in addition to their regular paychecks. Assuming that exerting effort is costly, we may view E as the risky strategy and L as the safe strategy. The safe payoff is the value of the regular paycheck of a player who exerts no effort and receives no bonus. However, due to the bonus condition, all $t_i = n$ and $\kappa(\overline{E}) = n$. So in this game $\overline{E}$ is a fragile equilibrium; $\rho(\overline{E}) = 0$ because if one worker fails to exert effort, then it is a best response of the others to not exert effort.

The next family of games provide natural interpretations for the resilience index, ρ.

4.3. **Graph matching games.** These games describe a large variety of situations in which the players' payoffs depend on the number of their neighbors that their choice matches. In communication games, players may wish to match the language choice of their neighbors, while in political games they may wish to match the political system advocated by their neighbors, as is the case in many other interactions in which players wish to match the standards, conventions, and mores of their neighbors (See Jackson (2008)).

Formally, a graph matching game Γ is described as three tuple $\Gamma = (N, A, C)$, in which $N = \{1, 2, \dots, n\}$ is a set of players, or nodes in the underlying directed graph; $A \subseteq \{(j, i) \in N \times N : j \neq i\}$ is a set of *arcs* that describe payoff implications (i.e., $(j, i) \in A$ indicates that player j's strategy affects player i's payoff); and C is a set of *choices*, available to all the players. In the illustrations that follow we assume that C contains at least two distinct elements denoted by H and S. The set $N_i \equiv \{j \in N : (j, i) \in A\}$ denotes the set of *neighbors* of player i. For every profile of choices θ (with each $\theta_i \in C$), $u_i(\theta) = |\{j \in N_i : \theta_j = \theta_i\}|$ (if $N_i = \emptyset$, $u_i(\theta) \equiv 0$).

Next, we follow an indirect method to compute the resilience $\rho(\overline{H})$ of the agreement profile $\overline{H}$ in which all the players choose H. First, for every player i we define $s_i = (|N_i|/2) + 1$, if $|N|_i$ is even; $s_i = (|N_i/2|) + 1/2$, if $|N_i|$ is odd; and $s_i \equiv n$, if $N_i = \emptyset$. Notice that s_i is the minimal number of neighbors of player i who can strictly incentivize player i to defect from $\overline{H}$. That is: at any profile θ in which s_i of i's neighbors choose S, S is a best response of player i. We consider a player v to be most vulnerable, if $s_v = \min_i s_i$.

Proposition 4. *The resilience of* $\overline{H}$ *is given by* $\rho(\overline{H}) = \min_i s_i - 1$.

Proof. First notice that by the definition of s_i, if $d \leq \min_i s_i - 1$, then at any profile θ with d or fewer $\overline{H}$-defectors, H is a best response for every player. Thus by Remark 1, $\min_i s_i - 1 \leq \rho(\overline{H})$. To see the converse, consider a most vulnerable player v and any profile η in which H is chosen by all the players except for s_v of v's neighbors who choose S. By the definition of s_i it is clear that H is not ν's best response to η; thus, $\min_i s_i > \rho(\overline{H})$. ∎

From the proposition above we conclude that for any graph Γ, $\kappa(\overline{H}, \Gamma) = n + 1 - \min_{i \in N} s_i$.

4.4. **Decentralization, operations management, and political interactions.** In this section, we use critical-mass analysis in graph-matching games to provide simple explanations for well-known issues of stability in economics, political science and operations management. More specifically, for the agreement profile $\overline{H}$, we discuss the critical mass and resilience values, $\kappa(\overline{H},\Gamma)$ and $\rho(\overline{H},\Gamma)$, for the five well-known n-node graphs Γ listed below. Then, we discuss the implications of these values to specific issues of stability.

(1) In the fully disconnected graph $\digamma$ in which $N_i=\emptyset$ for every player i: $\kappa(\overline{H},\digamma)=1$ and $\rho(\overline{H},\digamma)=n-1$, i.e., $\overline{H}$ is a dominant strategy equilibrium.

(2) In the complete graph Δ in which $N_i=N_{-i}$ for every player i: $\rho(\overline{H},\Delta)\approx n/2\approx\kappa(\overline{H},\Delta)$, with the precise values depending on whether n is even or odd.

(3) In the star-shaped graph Λ in which $N_1=\emptyset$, and $N_i=\{1\}$ for $i=2,\ldots,n$: $\kappa(\overline{H},\Lambda)=n$ and $\rho(\overline{H},\Lambda)=0$, i.e. $\overline{H}$ is a fragile equilibrium.

(4) In the three-stars graph 3Λ in which $N_i=\emptyset$ for $i=1,2,3$, and $N_i=\{1,2,3\}$ for $i=4,\ldots,n$: $\kappa(\overline{H},3\Lambda)=n-1$ and $\rho(\overline{H},3\Lambda)=1$, i.e., $\overline{H}$ is not as fragile when $\Gamma=3\Lambda$ as it is when $\Gamma=\Lambda$.

(5) In the linear graph Σ, in which $N_1=\emptyset$; and $N_i=\{i-1\}$ for $i=2,\ldots,n$: $\kappa(\overline{H},\Sigma)=n$ and $\rho(\overline{H},\Sigma)=0$, i.e., $\overline{H}$ is a fragile equilibrium.

In item 1 above, it is not surprising that for players who are fully indifferent to each other $\overline{H}$ (or any other profile) is a dominant strategy equilibrium.

In item 3 above, the fragility expressed by $\kappa(\overline{H},\Lambda)=n$ and $\rho(\overline{H},\Lambda)=0$ was illustrated in our chips game, see Example 2 in the introduction. In addition to the chips situation presented there, this type of fragility is present in other strategic interactions in which all the players' choices are guided by one **central decision maker.** Examples include other decisions made in centralized production and distribution systems, in social systems controlled by a dictator, in monetary systems controlled by one central bank, etc.

Item 2 above, illustrates the higher stability attained at fully unguided **decentralized decisions**, as modeled by the fully connected graph $\Gamma=\Delta$ with $\kappa(\overline{H},\Delta)\approx n/2$

and $\rho(\overline{H}, \Delta) \approx n/2$; in comparison with the lower stability of centrally guided decisions, as modeled by the star shaped graph $\Gamma = \Lambda$, with $\kappa(\overline{H}, \Lambda) = n$ and $\rho(\overline{H}, \Lambda) = 0$. This difference is clearly illustrated language choice game below.

Example 5. ***Language choice****: Each person in a country of 300 milion people has to choose a language from a set of languages that includes H. Consider two countries: a centralized one C, and a decentralized one D. In C, every player want to choose the language that matches the choice of one specific player, say player* #1*; whereas in D, every player wants to choose a language that maximized the number of matches with any people in the country.*

Consider the resilience of the profile $\overline{H}$, in which everybody chooses the language H. A defection by a single player, in particular #1, is of concern to the H choosers in C. On the other hand H choosers in country D have only small concerns about defections. As long as the number of defectors is smaller than 150 million players, the choice of H is optimal.

Item 4 above illustrates that player **replications** may serve as a mean to improve the resilience of agreement profiles in guided strategic interaction. For example, in the chip production game, if players $1, 2, 3$ are chip producers and the remaining $n-3$ players are chip users (each wishing to match the majority of producers), then the replication of producers raises the resilience level from $\rho(\overline{H}, \Lambda) = 0$ to $\rho(\overline{H}, 3\Lambda) = 1$. Similar increases in resilience are obtained if a decision guided by one star player is replaced by a decision guided by a group. For example, we may use the graph 3Λ to describe a political environment in which a dictator was replaced by a three-member politburo: players $\{1, 2, 3\}$ are the **politburo** members, and each of the subordinate players, $4, \ldots, n$, wants to match the choice of a maximal number of politburo members.

Item 5 above illustrates another type of fragile equilibria in production-distribution games defined by a linear graph Σ. **Just-in-time production** and **supply-chain** games may be described by such graph-matching games.

4.5. **Equilibrium Adoption and Sustainability.** The uniformity property in the definition of m-incentive-compatibility, implies that any group L of $m-1$ π-players (strongly) incentivizes all the group outsiders to play their π strategies. More specifically, at every profile θ with a group L of $m-1$ π-players, π_j is a best response for any player $j \in L^c$ (no matter what the other $n-m-2$ outsiders in L^c play).

In particular, since $\kappa(\pi)$ is defined to be the $\min m$ for which π is m-incentive compatible, it is easy to see that $\kappa(\pi)-1$ is the smallest group size that can incentivize the adoption of π by all the group outsiders. Thus, adoption of profiles π with a small $\kappa(\pi)$ values is easy, since it can be accomplished by recruiting a small number of players ($\kappa(\pi)-1$) to play their π strategies.[1]

The resilience index, $\rho(\pi)$ $(= n - \kappa(\pi))$ provides useful information about the resilience of profiles π to defection. Recall that a defection by any group of up to $\rho(\pi)$ players is not sufficient to incentivize any additional players to defect. Thus, to bring about defections from π with a large $\rho(\pi)$ values, one would have to recruit the participation of a large number of defectors.

The observations above help to explain the presence of two groups of equilibria often observed in large social systems:

Group 1: Equilibria π with small $\kappa(\pi)$ values and large $\rho(\pi)$ values. These equilibria are relatively easy to form and are highly stable against defections. Examples include the large number of academicians who subscribe to Zoom and teenagers who subscribers to Instagram.

[1] At any profile in which a group G of $\kappa-1$ players play π, it is a best response of all the G-outsiders to play π. But if G consists of κ π-players, π is best response of all the players, including both G-outsiders and G-insiders.

Group 2: Equilibria π with large $\kappa(\pi)$ values and large $\rho(\pi)$ values.[2] These equilibria may be difficult to adopt, because of their large $\kappa(\pi)$ values. But if they are formed by historical, legal, or other reasons, their large $\rho(\pi)$ values means that they are sustainable. Many equilibria that involve the social matching of conventions, standards, and mores are typical of this group. Examples include populations in whch all the people speak the same language, all use the same measurement system, all use the same currency, and all the men wear ties to job interviews.

Profiles π with small ρ values are rarely observed in social systems, apparently due to their low resilience against defection. An often discussed illustration is the honest ranking of candidates in the Beauty Contest Game, $\overline{H}$, (see Nagel (1995)}. In such games the only Nash equilibrium $\overline{H}$ is fragile (it is easy to show that $\rho(\overline{H}) < 1$) . Indeed, laboratory experiments show that players often do not play this unique Nash equilibrium; rather, they end up playing a nonequilibrium profile of strategies.

4.6. **Equilibrium implementation and switching.** The intermediate equilibrium concepts, $\mathcal{C}_2, \ldots, \mathcal{C}_{n-1}$, enable a broader set of social implementation concepts. We first illustrate this in a subtle multiperson Prisoners Dilemma game.

4.6.1. *Information revelation.*

Example 6. ***Conspiracy of Silence***. *Simultaneously, each of 100 conspirators may reveal a shared secret, R, or not reveal it (stay silent), S. If everybody chooses S, everybody is paid* 0*. But if some play R, then every R chooser is paid -1, and every S chooser is paid -3.*

Consider the mixed-strategy extension of the game, in which every conspirator may play the mixed strategies R^{λ}, for $0 \leq \lambda \leq 1$: choose R with probability λ and choose S with probability $1-\lambda$. It is easy to see that $\overline{R}$, in which all reveal with probability

[2] Both κ and ρ may be large when n is large.

one, is the only nearly dominant strategy equilibrium, and that all the other Nash equilibria are fragile. These fragile equilibria include $\overline{S}$, in which everybody stays silent with probability 1. We note that the fragility of $\overline{S}$ holds, despite the fact that in the game-theoretic language of equilibrium refinements, $\overline{S}$ is a strong and coalition-proof Nash equilibrium.

Indeed, as illustrated in movies about crime syndicates, syndicates are typically concerned about the play of the nearly dominant strategy equilibrium $\overline{R}$. To reduce the likelihood of the play of $\overline{R}$, syndicates change the game: they drastically "lower the payoffs" of R choosers, so players cannot be incentivized to choose R.

4.6.2. *Nearly dominant strategy implementation.* Consider n (≥ 3) bidders about to bid on a government oilfield of \$$Q$ (> 1) net-worth of oil. Each of the bidders knows the value Q, but the government does not. Conducting the auction in values rounded down to billions of dollars, the net-value of the field is q^* billion dollars. In a first-price sealed-bid auction (FPSBA) every bidder $i = 1, 2, \ldots, n$ is asked to submit an integer bid $b_i = 0, 1, 2, \ldots$. The government will identify the highest submitted bid, b^*; choose at random one of the highest bidders j^*, i.e. $b_{j^*} = b^*$; and award j^* the exclusive use of the oilfield in exchange for the payment of b^* billion dollars.

Consider the profile of honest bids, $\overline{q}^*$, in which every bidder bids q^* = the largest rounded-down integer value of Q. How reliable is the honest-bid profile $\overline{q}^*$? It is easy to see that $\overline{q}^*$ is not a dominant strategy equilibrium, yet $\overline{q}^*$ seems to be a highly reliable equilibrium since it is a nearly dominant strategy equilibrium, i.e., $\kappa(\overline{q}^*) = 2$. Thus, $\overline{q}^*$ is incentive compatible for any two or more bidders. In other words, the belief that at least one opponent bids q^* makes the bid q^* a dominant strategy for all the others.

4.6.3. *Eliaz Implementation.* The results of Eliaz and the current paper reinforce each other. From the mathematical definitions, a player who participate in an implementation scheme is faulty in Eliaz terminology iff the player is an (unrestricted) defector from the implementor's equilibrium in the terminology of the current paper. So based on the properties of the resilience index discussed in the current paper, for an implementing equilibrium of high resilience, the implementor may count on the incentives of the non-faulty players to play the equilibrium despite the presence of the faulty players.

Consider the first-price sealed-bid auction above, in which the honest bid equilibrium $\overline{q}^*$ has the critical mass $\kappa(\overline{q}^*) = 2$. This means that the resilience index $\rho(\overline{q}^*) = n - 2$. For larger values of n, the relatively large ρ value gives rise to another appealing property of the FPSBA above. Suppose, as assumed by Eliaz (2002), that the number of faulty bidders (i.e., ones with imperfect information about the value of Q or ones who use incorrect computation methods) is at most $n - 2$. Because $n - 2 \leq \rho(\overline{q}^*)$, we conclude that, regardless of the presence of the faulty bidders, it is still a best response for any nonfaulty bidder to bid q^*. Thus, the FPSBA should yield the government a payoff of at least \$$q^*$.

4.6.4. *Equilibrium switch in a ride-share game.* Each of 12 passengers has to choose one of two options: sign up for a private taxi ride (T) at the cost of \$100, or sign up for a shared van ride (V) that can comfortably accommodate any number of them. The cost of the van, \$180, will be shared equally by all the van choosers.

Consider an existing NE, $\overline{T}$, in which all 12 passengers choose private taxis; and the potentially competing equilibrium $\overline{V}$, in which all 12 choose the van. It is easy to see that $\overline{T}$ is a fragile equilibrium, with $\kappa(\overline{T}) = 12$. On the other hand, $\overline{V}$ is a nearly dominant-strategy equilibrium, $\kappa(\overline{V}) = 2$; thus, one player signing up for the van makes $\overline{V}$ a dominant strategy for all the others. This low $\kappa(\overline{V})$ value leads to highly plausible switching mechanisms, as illustrated below.

A van company voucher scheme: The van company offers one passenger - say number 3 - a voucher that will cover any of her van costs in excess of \$50. Passenger 3 accepts the voucher and signs up for the van. With one passenger choosing V, it is a dominant strategy for each of the 11 remaining passengers to also choose it. Assuming that they all follow their conditionally dominant strategy, the van company succeeds in switching everybody to the van. Each van chooser pays 180/12=\$15, and the voucher is actually not used.

Passenger-initiated schemes: Figuring out the logic of the voucher scheme above, one strategically minded passenger signs on to the van without the voucher, the rest of the passengers follow, and everybody pays \$15 for the van ride.

The easy adoptions of $\overline{V}$ above, which are due the low value $\kappa(\overline{V}) = 2$, are complemented by $\overline{V}$'s high resilience to defection, which are due the high resilience index $\rho(\overline{V}) = 10$. In particular, no players gain by defecting from the van, unless all the 11 van choosers co-defect with them.

To emphasize the importance of the low $\kappa(\overline{V})$ value, consider a modified version of the Ride Share game above, in which the cost of the van is modified to be \$1000 instead of \$180. It is easy to see that now $\kappa(\overline{V}) = 1000/100 = 10$, and $\kappa(\overline{V}) - 1 = 9$. Thus, in an appropriately modified voucher scheme, the van company would have to offer 9 passengers vouchers that cover any van costs, in excess of, say \$90. This modified voucher scheme is based on the mutual trust that all 9 passengers would accept the voucher and sign up for the van, and that the remaining 3 passengers would be aware of this and sign up for the van too. If there is sufficient mutual trust to make this work, then every passenger would end up paying 1000/12=\$83.33, and the \$90 vouchers would not be used. But the required high level of mutual trust makes the voucher scheme significantly less plausible now.

The passenger-initiated schemes are also significantly less plausible in the modified game. Instead of one strategically minded passenger initiating a switch to the van (without a voucher), in the modified scenario 9 such passengers must choose to participate in a mutually coordinated incentive-compatible van-choosing scheme based on mutual trust among all 9, and the remaining passengers have to trust that 9 strategically minded ones did so.

Cardinal computations of critical mass. The comparison of equilibria when the cost of the van is changed from \$1000 to \$180, with the corresponding change of the equilibria from $\overline{V}_{\$1000} \in \mathcal{C}_{10}$ to $\overline{V}_{\$180} \in \mathcal{C}_2$ brings up an important observation about the $\mathcal{C}_k$ equilibria and the axioms that characterize them. A reader who views the reduction of the van cost as an increase in the payoff functions of all the van choosers can take a shortcut through the payoff monotonicity axiom to conclude that since $\overline{V}_{\$1000} \in \mathcal{C}_{10}$, $\overline{V}_{\$180}$ should remain in $\mathcal{C}_{10}$. While this is correct ($\mathcal{C}_2 \subseteq \mathcal{C}_{10}$), the actual cardinal computations done above lead to the stronger conclusion $\overline{V}_{\$180} \in \mathcal{C}_2$. As computed above, the number of passengers needed to justify the choice of the van is actually $\kappa(\overline{V}_{\$180}) = 2$.

4.6.5. *The Swedish equilibrium switch, Dagen H.* In contrast to the easy switch of 12 passengers from the taxi to the van in the example above, explained by the indices κ and ρ, these same indices provide an explanation for why switching a large number of drivers n from all driving on the left side of the road ($\bar{L}$) to all driving on the right side ($\bar{R}$) is difficult. Indeed, it took a great effort by the Swedish government to bring about this switch, which was implemented in Sweden on *Dagen H*, September 3, 1967.

Consider cases in which the values $\kappa(\bar{L})$, $\rho(\bar{L})$, and of $\kappa(\bar{R})$, $\rho(\bar{R})$, are all large, for example one half of the number of Swedish drivers. If players think of the switch from $\bar{L}$ to $\bar{R}$ as being made in two steps, then the reasoning outlined in Section 4.5

suggests that each step would be difficult: In the first step, one has to convince the drivers to stop following their L strategy, which is difficult due to the large resilience value, $\rho(\bar{L})$. In the second step, convincing them to adopt the R strategy is also difficult due to the large critical mass value, $\kappa(\bar{R})$.

The Dagen H project in Sweden involved a large and extensive government campaign that relied on informational, educational, and legal instruments.

5. Conclusions and future research

5.1. **A perspective on equilibrium concepts.** Equilibrium concepts for n-person strategic games are some of the oldest and most studied topics in game theory. Discussions on this topic deal mostly with n-person generalizations of optimization by a single decision maker. The most immediate generalization is the **dominant strategy equilibrium** (DSE). Importantly, DSE is immune to *faulty opponents' strategies*, i.e., a player's dominant strategy is an optimal choice no matter what strategies are chosen by the opponents. Unfortunately DSE fail to exist even for most elementary games. The nonexistence of DSE motivated the development of other, less ambitious equilibrium concepts.

The **maxmin equilibrium** of von Neumann overcomes the non-existence difficulty of DSE, but it suffers two shortcomings. First and foremost, this equilibrium concept is restricted to two-person zero-sum games. But a second serious shortcoming regards faulty opponents' strategies: When player i plays a maxmin strategy against a nonmaxmin opponent j, i's strategy is most likely a suboptimal response to j's. This is so despite the fact that i's payoff is bounded below by the payoff i would have received against a maxmin playing opponent.

The **Nash Equilibrium** concept overcomes the first shortcoming as illustrated by the *Nash existence theorem*: under relatively mild assumptions on the payoff functions associated with the mixed-strategy extension of the game, Nash equilibria (in mixed strategies) exist. However, the concern about faulty opponents' behavior becomes even more sever. A Nash player's strategy can be unboundedly suboptimal against faulty opponents.

The **critical mass equilibria** studied in this paper, $\mathcal{C}_m$, constitute a decreasing hierarchical progression of equilibrium concepts, arranged by increasing concerns about faulty opponents behavior. The strategy of a $\mathcal{C}_m$-player remains optimal against faulty opponents as long as (the number of faulty opponents) $\leq n - m$. Thus, for $m = 1$, a $\mathcal{C}_1$-player (i.e., a dominant strategy chooser) is unconcerned about the strategies chosen by the opponents. At the other end of the progression, a $\mathcal{C}_n$-strategy chooser (i.e., a chooser of an arbitrary NE strategy) can find that the chosen strategy is unboundedly suboptimal against faulty opponents.

Researchers who wish to use an equilibrium concept that is optimal against any faulty opponents must assume some bounds on the possible faulty behavior; otherwise, they would be led back to the use of dominant strategy equilibrium and its nonexistence issues. One bound used in the equilibrium refinement literature is the assumption that faulty opponents choices can only be made with "trembling hands," and thus the faulty choices must be arbitrarily close to the perfect equilibrium choices. Despite such minimal faulty behavior, this approach still leads to significant improvement in the analysis of strategic interactions. The critical mass equilibrium in this paper assumes that faulty opponents may play any game-feasible strategies (even ones that are far from the equilibrium), but it places a bound on the possible number of such faulty opponents. The critical-mass view leads to an equilibrium ranking of all strategic equilibria. The examples in this paper deal with strategic interactions studied in a variety of different disciplines. In these examples, the critical-mass ranking is significantly stronger than that of the Nash equilibrium and its refinements.

5.2. **Summary and future research.** Critical mass analysis addresses a variety of strategic issues that are hard to explain or even identify by means of the commonly used equilibrium concepts and their refinements. By bringing such issues down to the fundamental equilibrium concepts, the intermediate critical mass equilibria offer a unified and simple way of explaining them. The definitions and ranking of these equilibria follow from three axioms: two are standard axioms commonly used in decision theory and other conventional areas of economics and one that addresses concerns about faulty opponents.

In this paper, the critical mass equilibrium concepts, that follow from the three axioms referred to above, are defined and ranked by the critical mass index of stability κ. In exactly the same manner, one may define and rank the critical mass equilibrium concepts by any index of stability σ that preserves (or reverses) the order over strategy profiles imposed by κ (e.g., $\sigma = 100\kappa/n$ or $\sigma = 1/\kappa$). Such indices may be useful in more general future research, such as situations in which the number of players n is not fixed.

Regardless of the index we use to describe the critical mass equilibria, a mathematical question of interest is a *generalization of the Nash existence theorem* from $\mathcal{C}_n$, to $\mathcal{C}_m$'s with $m < n$. What are natural conditions on the utility functions u_i (beyond continuity and convexity) of a game u that would imply that $\mathcal{C}_m(u) \neq \emptyset$? Given the substantial mathematical arguments presented by John Nash to establish the nonemptyness $\mathcal{C}_n(u) \neq \emptyset$ for a broad general class of payoff function u's, this seems to be a difficult problem that should be left for future mathematical research.

The difficulty in proving the existence of the $\mathcal{C}_m$ equilibria, seems to be in contrast to the prefect and proper equilibrium concepts of Selten and Myerson, see Fudenberg and Tirole (1991, pp. 355-7). A key difference is that perfect and proper equilibrium require stability against infinitesimal deviation from equilibrium by any of the players. As such, their existence can be established by subtle refinements of the fixed point arguments of Nash. The stability of the $\mathcal{C}_m$ equilibrium, on the other hand, requires stability against unbounded deviations from the equilibrium by a bounded number of players. This expands the applicability of the $\mathcal{C}_m$'s but makes the analysis of existence more difficult.

This paper studies the critical mass index in strategic (static normal form) games and leaves out important models designed to deal with long run play, dynamic play, Markovian play, Bayesian analysis, and so forth. An important first step to deal with these more advanced issues is the extension of the critical mass analysis, presented here, to extensive form games. In a similar manner to the development of perfect equilibrium by Selten, one may define the critical mass of an equilibrium in an extensive game to be the critical mass of the equilibrium in the associated static strategic game. However, in this association one encounters at least two competing options: Associate to the extensive game the "standard normal form" or the "agent normal form." This is especially important for the notion of critical mass, which counts the number of defections from equilibrium. For example, the standard normal form game would allow cumulative addition in the counted number of defection as the game progresses, increasing the fragility of long-run equilibria. On the other hand, the agent normal form would restart the count of number of defections from equilibrium in every new stage of the game, which may be more proper for Markovian long run analysis. We feel that at this initial stage, it is best to study the issues above in specific examples that may lead to a later general model.

A related but different direction of research would be a study of bolder alternatives of equilibrium stability indices that are more refined and go beyond the order imposed by κ. For example, one may assign different weights to different players, which may be useful for modeling games such as the heterogeneous investor implementation of Halac, Kremer, and Winter (2020). Such equilibrium concepts would have to violate some of the axioms used in this paper, but they may still be useful, similar to concepts in the equilibrium refinement literature.

Acknowledgments. The authors acknowledge helpful discussions with Nemanja Antic, Robert Aumann, Vince Crawford, Kfir Eliaz, Drew Fudenberg, Ronen Gradwohl, Matthew Jackson, Julien Manili, Eric Maskin, Joshua Mollner, Harry Pei, Larry Samuelson, James Schummer, Eran Shmaya, Jörgen Weibull, Peyton Young, Ro'i Zultan; and seminar participants at the Universities of the Basque Country, Oxford, Tel Aviv, Yale, Stanford, Berkeley, Stony Brook, Bar Ilan, the Technion, the Hebrew University, Brown, Rochester, Montreal, Northwestern, as well as Microsoft Research. The authors thank Vince Crawford for his suggestion of the term *fragile equilibrium*.

References

Abraham, I., D. Dolev, R. Gonen, and J. Halpern (2006), "Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation." In *Proceedings of the. 25th ACM Symposium on Principles of Distributed Computing*, 53–62.

Aumann, R. (1959), "Acceptable points in general cooperative n-player games," in Tucker, A. and Luce, R., Eds., *Contributions to the Theory of Games IV*, Princeton Univ. Press, Princeton.

Bernheim, B. D., P. Bezalel, and M. Whinston (1987), "Coalition-Proof Nash Equilibria I. Concepts," *Journal of Economic Theory*, Volume 42, Issue 1, 1–12.

Crawford, V. P. (1995), "Adaptive Dynamics in Coordination Games," *Econometrica* 63, No. 1, 103–143.

Crawford, V. P. (2001), "Learning Dynamics, Lock-in, and Equilibrium Selection in Experimental Coordination Games," in Ugo Pagano and Antonio Nicita, editors, *The Evolution of Economic Diversity (papers from Workshop X, International School of Economic Research, University of Siena)*, London and New York: Routledge, 2001, 133-163.

Deepanshu, V., and R. Berry (2020), "Fault Tolerant Equilibria in Anonymous Games: best response correspondences and fixed points," ResearchGate `https://www.researchgate.net/publication/341115293`.

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

Fudenberg, D., and J. Tirole (1991), *Game Theory*. MIT Press, Cambridge MA.

Goldreich, O., S. Goldwasser, and N. Linial (1998), "Fault-tolerant computation in the full information model," *SIAM Journal on Computing* 27.2: 506–544.

Gradwohl, R. and O. Reingold (2014), "Fault tolerance in large games," *Games and Economic Behavior*, 86, 438–457.

Halac, M., I. Kremer, and E. Winter (2020), "Raising Capital from Heterogeneous Investors," *American Economic Review*, 110(3), 889–921.

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

Jackson, M. O. (2008) *Social and Economic Networks*. Princeton Univ. Press, Princeton.

Kalai, E. (2019), "Viable Nash Equilibria: Formation and Sustainability." `https://www.researchgate.net/publication/335168588`

Kalai, A. T. and E. Kalai (2022), "Best-response reasoning leads to critical-mass equilibria." `https://www.researchgate.net/publication/356911802`

Kandori, M., G. Mailath, and R. Rob (1993), "Learning, Mutation, and Long Run Equilibria in Games." *Econometrica*, 61, No. 1, 29–56.

Kaneko, M. (1995), “Aximatic Considerations of Nash Equilibrium,” January 1995, Bulletin of the section of Logic 24(1): 6-12

Kim, D. G., D. Min and J. Wooders (2022) “Viable Nash Equilibria: an Experiment,” Discussion paper, Division of Social Science, New York University, Abu Dhabi.

Myerson, R. B. (1978), “Refinements of the Nash equilibrium concept,” *International Journal of Game Theory*, No. 7, 73â€“-80.

Myerson, R. B. (1997), *Game Theory: Analysis of Conflict.* Harvard University Press, Cambridge.

Nagel, R. (1995), “Unraveling in guessing games: An experimental study,” *The American Economic Review*, 85, No. 5, 1313–1326.

Savage, L. J. (1954), *The foundations of statistics.* John Wiley & Sons Inc., New York.

Salonen, H. (1992), “An axiomatic analysis of the Nash equilibrium concept,” *Theory and Decision* 33, 177-189.

Selten, R. (1975), “Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games,” *International Journal of Game Theory* (4), 22–55.

Schelling, T. (1973), “ Hockey Helmets, Concealed Weapons, and Daylight Saving, a Study of Binary Choices With Externalities,” *Journal of Conflict Resolution*, 17, No. 3: 381-4-28.

Skyrms, B. (2001), “The Stag Hunt,” *Proceedings and Addresses of the American Philosophical Association* (2), 31–41.

Young, H. P. (1993), “The Evolution of Conventions.” *Econometrica*, 61, No. 1, 57–84.