Matching with Strategic Consistency*

Marzena Rostek† and Nathan Yoder‡

March 20, 2026

Abstract

In many environments, agents form agreements that have externalities or are multilateral, and may view some agreements as substitutable and others as complementary. This paper presents an approach that ensures the existence of stable outcomes in any environment, including those with arbitrary externalities, preferences, and market structures. It does so by endogenizing the agents' choice functions while employing the standard stability concept. Instead of assuming that each agent chooses their favorite set of contracts, we require agents to choose optimally given correct beliefs about the choices of others, and show that stable outcomes are uniquely pinned down by those beliefs.

Keywords: Externalities, multilateral matching, matching with contracts, stability

* This paper subsumes parts of "Matching with Multilateral Contracts". The authors are grateful to numerous colleagues for their helpful comments and suggestions. We also thank conference attendees at the North American Summer Meetings of the Econometric Society at UCLA, the ASSA in New Orleans, 33rd Stony Brook International Conference on Game Theory, the 14th Conference on Economic Design at Essex, and seminar participants at Virtual Seminars in Economic Theory (VSET), Brown University, and University of Oregon. This material is based upon work supported by the National Science Foundation under Grant No. SES-1357758. Yoder acknowledges summer support from the University of Georgia through a Terry-Sanford Research Award.

† University of Wisconsin-Madison, Department of Economics; E-mail: mrostek@ssc.wisc.edu.

‡ University of Georgia, Terry College of Business, John Munro Godfrey, Sr. Department of Economics; E-mail: nathan.yoder@uga.edu.

1 Introduction

Matching theory has facilitated the study of many applications where agents negotiate agreements with one another. In particular, the literature has extensively explored settings where these agreements (sometimes referred to as *contracts*) are substitutable, bilateral, and do not have externalities. In these environments, the literature has established the general existence of *stable outcomes* — sets of agreements that are robust to both individual deviations to remove agreements and joint deviations to form new ones — and provided ways to find them (e.g., the seminal work of Gale and Shapley (1962), Kelso and Crawford (1982), Hatfield and Milgrom (2005), and Hatfield et al. (2013)).

But in many environments, agents may form agreements that have externalities, are not substitutable, or are multilateral. Agreements among competing firms to merge or engage in collusion may exert externalities on other market participants, influencing their incentives to enter into alternative agreements. These agreements might involve more than two firms. And depending on the market structure, some could be complementary, while others are substitutable. Analogous features characterize international treaties, legislative negotiations to pass a bill, and agreements to add a healthcare provider to an insurance network, among others. The prevalence of these features has created demand in applied research for matching-theoretic tools that are capable of accommodating them (e.g., Agarwal et al. (2021)).

Accommodating any of these features in general matching environments has been challenging, because each prevents standard approaches from guaranteeing the existence of stable outcomes.[1] This paper introduces an approach that allows matching-theoretic stability to be applied in any setting, including those with general externalities, arbitrary preferences and market structures, and multilateral agreements. Our key observation is that the challenges presented by these features can each be attributed to an implicit assumption about the way that choices are derived. Specifically, when agents choose from a set of available agreements, they select their *favorite* subset, thus behaving as if each available contract will go into effect if they choose it. However, for contracts to go into effect, they must also be chosen by other agents. We show that when agents take others' choices into account — i.e., they choose optimally given beliefs about others' choices, and those beliefs are correct at every set of available agreements and consistent (in a sense akin to independence of irrelevant alternatives) across sets of available agreements — stable outcomes always exist (Theorems

[1] When both complementarity and substitutability are present in the same environment, the existence of stable outcomes is generally not guaranteed (Hatfield and Kojima (2008)). Moreover, standard existence results do not always apply in the presence of externalities (Sasaki and Toda (1996)) or in the absence of a key assumption on market structure (acyclicity) that is incompatible with multilateral agreements (Gale and Shapley (1962); Hatfield and Kominers (2012)).

1 and 2). Thus, we ensure existence not by modifying the usual definition of stability, but by endogenizing the agents' choice functions and the beliefs that generate them. We call such a profile of choice functions and beliefs *strategically consistent*.

This result does not require conditions on preferences (e.g., (full) substitutability or complementarity), market structure (e.g., acyclic trading networks), or the agents that agreements can involve (e.g., bilateral agreements) or affect (e.g., no externalities). The literature has demonstrated that these conditions ensure that stable outcomes can be represented as fixed points of monotone operators; by Tarski's theorem, such fixed points always exist.[2] Theorem 2 instead constructs fixed points in profiles of *choice functions and beliefs* for all agents, without relying on a monotonicity condition or a fixed point theorem. Each such profile of choice functions then pins down a unique stable outcome (Theorem 1).

Given the fixed point relationship between optimal choices and correct beliefs, there may be many outcomes that are stable for *some* strategically consistent profile of choice functions and beliefs. In particular, restricting attention to such outcomes does not allow us to rule out any outcomes that could not be ruled out by simply requiring individual rationality (i.e., that individuals cannot profitably drop some of their contracts) (Proposition 1). Thus, if we wish to use stability in our framework as a tool to rule out additional outcomes, we must focus on *specific* profiles of choice functions and beliefs.

One way to do this is by placing additional structure on agents' beliefs about others' choices using a refinement. We consider three of these in this paper.[3] The first rules out failures to coordinate on blocks that are individually rational and Pareto-superior. As it turns out, one can construct such profiles by solving a set of constrained social planner's problems (Theorem 3). This provides one direction of a "welfare theorem" for strategic consistency (Theorem 4), which parallels welfare theorem results in the matching literature with transferable utility (e.g., Hatfield et al. (2013)). But since it decentralizes efficient outcomes with a profile of (common) *beliefs* rather than *prices*, Theorem 4 applies even to nontransferable utility settings, and without the conditions on preferences or market structure that are, in general, necessary for competitive equilibrium prices to exist.

The other two refinements we consider capture forward induction reasoning by the agents when they evaluate the credibility of proposed deviations. *Weak forward induction* requires that when a proposal to add contracts is credible, in the sense that no agent can benefit from rejecting any of the proposed contracts, either directly (by getting a higher payoff) or indirectly (by leading to new opportunities for further deviations), each agent should believe

[2]Since the set of stable outcomes is discrete, fixed points are not guaranteed more generally.

[3]We also show that the usual approach to stability is a refinement as well (Theorem 7), though it may eliminate *all* outcomes.

that the others will go along with it. This rules out all but the maximal outcomes predicted by individual rationality alone (Theorem 5). When we apply this reasoning to *all* blocking proposals, Theorem 6 gives a condition on the set of individually rational outcomes which ensures that the resulting refinement (which we call simply *forward induction*) still predicts at least one stable outcome.

Alternatively, if we observe an outcome empirically, we can use that outcome to select beliefs — and thus a stable outcome — in a counterfactual scenario (e.g., when the government levies a tax or a regulator disallows a contract). We explore this avenue in a companion paper, Rostek and Yoder (2025), where we show how one can use the beliefs in our framework to perform the same kind of counterfactual analysis that bargaining weights in Nash bargaining (or Nash-in-Nash bargaining) facilitate in empirical work (e.g., Ho and Lee (2017)).[4]

Because our approach allows general externalities, market structures, and preferences, and permits multilateral agreements, it enables the use of matching-theoretic stability in applications such as network formation, coalition formation, and bargaining with externalities. As we illustrate in Section 6, this allows one to make predictions in these environments that are robust to *arbitrary* deviations.

This is a larger set of deviations than is considered by existing tools used in these environments. Specifically, relative to perhaps the most common solution concept in network formation, *pairwise stability* (Jackson and Wolinsky (1996)), (matching-theoretic) stability permits agents to swap links (important when they are substitutes or with externalities) or add multiple links (significant when they are complements or with externalities).[5] And relative to *Nash-in-Nash bargaining* (Horn and Wolinsky (1988)), popular in applied work on environments with externalities, stability allows agents to simultaneously alter their agreements with multiple counterparties, endogenizes the agreements counterparties make, and allows agents to exclude counterparties by declining to make any agreements with them. The main obstacle to using stability in these contexts is that with the standard, nonstrategic approach to choice, these additional deviations create an existence problem that is not present with pairwise stability or the Nash-in-Nash solution. Rather than considering robustness to a smaller class of deviations, as those solution concepts do, strategic consistency sidesteps

[4]Observe that strategically consistent beliefs play the same role in our framework that bargaining weights play in Nash bargaining: Each profile of beliefs or bargaining weights predicts a *unique* outcome, but different profiles of beliefs or weights predict different outcomes. This allows us to use them to make predictions about a counterfactual environment in an analogous way: Specifically, the companion paper shows how we can recover parameterized beliefs from an observed stable outcome, and use the same parameters to pin down beliefs (and hence the stable outcome) following a change in the environment.

[5]Sadler (2023) also allows agents to swap links, rather than merely sever them, and establishes some of the classical matching-theoretic results in networks without externalities.

the existence problem by endogenously determining which deviations are relevant.

Similarly, in models where coalitions can form, our results ensure existence without imposing any restrictions on deviations allowed (e.g., only to subsets of coalitions) or coalitions that can form (e.g., partitions). In particular, because matching-theoretic stability requires outcomes to be robust to *arbitrary* deviations, its predictions are independent of assumptions about the specific ways that agents form coalitions.

Related Literature

Our paper relates to three strands of the matching literature. The first strand seeks to extend matching theory to accommodate agents' preferences over agreements that do not satisfy the classical substitutability condition. Several studies have demonstrated that the tools of matching theory can be applied to settings in which preferences satisfy more general forms of substitutability, such as full substitutability (Ostrovsky (2008); Hatfield et al. (2013); Fleiner et al. (2019)), or by applying substitutability under a basis change on the set of contracts, as in gross substitutes and complements (Sun and Yang (2006, 2009); Teytelboym (2014)). Another approach considers environments where all contracts are complementary rather than substitutable (Rostek and Yoder (2020)).[6] Other authors have shown that, instead of imposing restrictions on preferences, we can rely on conditions on the market structure (e.g., Bando and Hirai (2021)) or its size (e.g., Jagadeesan and Vocke (2021)), relax feasibility constraints (Nguyen and Vohra (2018)), or consider outcomes that are dynamically stable in markets with patient firms (Liu et al. (2023)). We show that when agents' choices are endogenized by requiring them to be optimal given correct beliefs about the choices of others, rather than being determined by a single-agent optimization problem, the existence of stable outcomes can be established for arbitrary preferences over agreements, market structures, and market sizes.

Second, our work also contributes to the literature on matching with externalities. One strand of this literature explores the externalities that arise in settings of applied interest, such as labor market matching with couples (e.g., Kojima et al. (2013)). Other studies consider general environments in matching markets with two sides (e.g., Bando (2012), Fisher and Hafalir (2016), Pycia and Yenmez (2023), and Liu et al. (2023)). Of these, our paper is closest to Pycia and Yenmez (2023), who introduce a matching with contracts framework in two-sided settings with a classical substitutability condition extended to allow for external-

[6] While our results in Rostek and Yoder (2020) also allow for multilateral contracts and externalities, these features are not a central focus there. With nontransferable utility, they do not create any additional challenges for the existence and characterization results from Rostek and Yoder (2020), precisely because complementarity ensures that whenever a block is relevant for stability, the implicit assumptions that nonstrategic agents make about other agents' choices turn out to be correct.

ities. Our results apply to environments where preferences may not satisfy substitutability and whose market structures may not be two-sided.[7]

Within the literature on two-sided matching markets with externalities, papers like Sasaki and Toda (1996), Hafalir (2008), and Saulle et al. (2025) consider agents who determine what to take as given about other agents' matchings through the use of an *estimation function*[8] — a concept akin to the beliefs considered in this paper, but which describes how agents who are *not* involved in a proposed deviation will *subsequently* respond to it, rather than how agents who *are* involved in the deviation will react to it being proposed.[9] The agent then evaluates potential partners by taking as given the *least preferred* outcome that is plausible according to her estimation function. While Sasaki and Toda (1996) take these estimation functions as a primitive of the model, Hafalir (2008) allows them to be determined based on a consistency condition: an agent's estimation function treats matchings as plausible if they are stable when the agent and her partner are removed from the market, given the set of plausible matchings. Saulle et al. (2025) weaken this endogeneity requirement (and thus strengthen the underlying stability concept) by instead requiring the absence of an individual or pairwise deviation that is profitable starting from *all* plausible matchings where the agents deviating have the same partners. In contrast, we require an agent's belief to be correct, i.e., match the choices made by other agents from the available set of contracts.[10]

Finally, our paper contributes to the literature on multilateral contracts. There is a large literature on the formation of coalitions or clubs; see, e.g., Pycia (2012); Ellickson et al. (1999). Hatfield and Kominers (2015) initiated the study of multilateral agreements in the matching with contracts framework. They examine settings with continuously divisible contracts and transferable utility, and leverage the concavity of agents' valuations to establish the existence of competitive equilibria (and thus, as they demonstrate, stable outcomes). As is standard in the literature, we work with environments where the set of contracts

[7] In Rostek and Yoder (2023), we focus on two-sided markets with externalities, and consider a weaker notion of strategic sophistication: Instead of requiring that agents have correct beliefs about *all* other agents' choices, we only require agents to have correct beliefs about the choices of others *on the same side of the market*. We show that the *standard substitutability* and *monotone externalities* conditions introduced by Pycia and Yenmez (2023) ensure the existence of profiles of choice functions and beliefs that satisfy this notion of strategic consistency, and hence stable outcomes. Intuitively, these conditions ensure that agents on the same side of the market can all form correct beliefs about each other's behavior.

[8] Saulle et al. (2025) instead use the term *system of conjectures*.

[9] While the beliefs considered in this paper specify the choices that an agent thinks others would make from a proposed set of contracts, estimation functions give a *set* of outcomes that an agent thinks are plausible, given the identity of the individual she is matched to.

[10] One approach to ruling out agents' disagreements about the outcomes of blocking proposals has been to strengthen the solution concept (e.g., setwise stability (Klaus and Walzl (2009))). Strategic consistency eliminates disagreements without strengthening the usual stability concept, while also ensuring existence even with externalities or non-substitutable preferences.

is discrete, rather than convex, and so the tools they use from convex analysis are not available. On the other hand, Bando and Hirai (2021) investigate a setting with a finite number of multilateral contracts, as we do. Unlike this paper or Hatfield and Kominers (2015), the authors use conditions on the market structure that guarantee the existence of stable outcomes, irrespective of agents' preferences.

Our paper also relates to several papers in the literature on matching with incomplete information that also explicitly incorporate agents' beliefs. In, e.g., Chakraborty et al. (2010), Liu et al. (2014), Liu (2020), and Liu (2022), agents form beliefs about other agents' privately observed *types*, and make choices given those beliefs. We do not consider incomplete information. Instead, beliefs in our paper are deterministic, and pertain to the contracts others will *choose* from each possible set of available contracts, rather than their types.[11]

The structure of the paper is as follows. Section 2 introduces the environment. Section 3 presents an example that illustrates the paper's main ideas, and provides our main existence results. Section 4 introduces refinements of strategic consistency, and provides our main characterization results. Section 5 provides additional discussion of conceptual issues. Section 6 discusses some of the novel applications of stability that are allowed by our approach.

2 Model

2.1 Setting

We work in a matching with contracts framework adapted to accommodate externalities and agreements among more than two agents.[12] Additionally, we do not assume a certain market structure (such as two-sidedness or acyclicity). Our model accommodates, for instance, network formation, many-to-many matching with contracts, and coalition formation.

There is a finite set I of agents and a finite set X of agreements, or *contracts*, that they can sign with one another. Each contract $x \in X$ requires the agreement of a set of agents $N(x) \subseteq I$ in order to be enacted. For sets of contracts $Y \subseteq X$, we write $N(Y) := \bigcup_{x \in Y} N(x)$. We assume that each contract involves at least two agents: For all x, $|N(x)| \geq 2$.[13] A contract

[11] In particular, the beliefs in our model are not equivalent to beliefs in the sense of Liu (2022) about the output of a correlation device: In cooperative games with incomplete information, Liu (2022, Theorems 1 and 6) shows that without payoff-relevant uncertainty, the presence of payoff-irrelevant signals cannot change the set of predictions consistent with stability. As we show, in matching models with complete information, considering beliefs to determine choice can alter the set of outcomes that are consistent with stability, in particular by making it nonempty (Theorem 1).

[12] Agreements between more than two agents cannot be represented by multiple independent bilateral agreements; see Example S.2 in the Online Appendix.

[13] For simplicity, we do not explicitly consider single-agent contracts. We can easily accommodate an

x is *multilateral* if $|N(x)| > 2$ and *bilateral* if $|N(x)| = 2$. For each agent $i \in I$, denote the set of contracts requiring i's agreement as $X_i := \{x \mid i \in N(x)\}$. In keeping with the literature, we say that X_i is the set of contracts that *name* i. Similarly, let $X_J := \bigcup_{i \in J} X_i$, let $X_{-i} := X \setminus X_i$, and for sets of contracts $Y \subseteq X$, write $Y_i := Y \cap X_i$ and $Y_{-i} := Y \cap X_{-i}$.

Each agent i has preferences over sets of signed contracts, or *outcomes*, which are represented by a utility function $u_i : 2^X \to \mathbb{R}_+$.[14] This allows for *externalities*: Agents' utility functions can depend on the presence of contracts that do not name them. In settings where it does not — i.e., when $u_i(Y \cup Z) = u_i(Y \cup Z')$ for each $Z, Z' \subseteq X_{-i}$ and $i \in I$ — we say that there are *no externalities*.

A *choice function* for agent i is a function $C_i : 2^{X_i} \times 2^{X_{-i}} \to 2^{X_i}$. Its arguments are the sets $Y_i \subseteq X_i$ and $Y_{-i} \subseteq X_{-i}$ of contracts that are *available* — those currently under discussion as part of a negotiation (over, e.g., a proposed new set of contracts) — to agent i and to agents other than i. When agent i's choice function is C_i, $C_i(Y_i|Y_{-i})$ gives the set of contracts that agent i chooses from the set of contracts Y_i available to him, given that the set of contracts available to other agents is Y_{-i}. Its second argument allows for the presence of externalities.

In Section 3, we describe two different ways in which these choice functions can be derived, given agents' preferences. In order to ensure that these endogenously derived choice functions are single-valued, we assume that agents' payoff functions have no indifferences, conditional on the set of contracts that do not name them: $u_i(Y \cup X') \neq u_i(Z \cup X')$ for each distinct $Y, Z \subseteq X_i$ and $X' \subseteq X_{-i}$.[15]

2.2 Stability

Our solution concept is the usual matching-theoretic definition of *stability*, generalized to our setting with multilateral contracts and externalities.[16]

extension where each agent i has a set of single-agent contracts (i.e., actions) A_i he can take by simply letting payoffs be given by $\hat{u}_i(Y) = \max_{S \subseteq A_i} u_i(S \cup Y)$.

[14] Throughout, we use 2^B to denote the power set of a set B.

[15] With nonstrategic choice, the presence of indifferences introduces additional wrinkles into the standard definition of stability, since it requires the consideration of choice *correspondences* (rather than functions). However, strategic consistency still pins down a profile of choice *functions*: Each agent must still form a belief about the available contracts that will be chosen by others, and correctness requires those beliefs to match the actual choices that those other agents make (rather than one of several choices returned by their choice correspondences). Consequently, our main results go through unchanged. Nevertheless, we rule out indifferences in our main analysis in order to avoid the discussion of choice correspondences that they necessitate with nonstrategic choice, which is not our main focus. We discuss these issues further in Section S.3 of the Online Appendix.

[16] In particular, our solution concept coincides with those of Gale and Shapley (1962) (one-to-one matching), Hatfield and Milgrom (2005) (many-to-one matching with contracts), and Hatfield and Kominers (2012) (matching on networks) in the settings they consider.

Definition (Stability). Given choice functions $\{C_i\}_{i\in I}$, a set of contracts $Y \subseteq X$ is *stable* if it is

i. *Individually rational*: $Y_i = C_i(Y_i|Y_{-i})$ for all $i \in I$.

ii. *Unblocked*: There does not exist a nonempty $Z \subseteq (X \setminus Y)$ such that for all $i \in N(Z)$, we have $Z_i \subseteq C_i((Z \cup Y)_i|(Z \cup Y)_{-i})$.

In words, a set of contracts Y is stable if (i) when Y is the set of available contracts, no one rejects any contracts from it (individual rationality), and (ii) no group of agents can propose to change the set of contracts in place by adding a new set of contracts Z, or *block*, that they are each willing to choose when made available (i.e., discussed in a negotiation) alongside Y.[17]

We accommodate externalities by allowing agents who participate in a block to take into account the contracts available to the agents they negotiate with: the second argument of the choice function in (ii) includes both the existing contracts Y_{-i} and blocking contracts Z_{-i} that do not name agent i.[18]

3 Strategic Consistency

This section presents the paper's main idea, which stems from a simple yet crucial observation about the relationship between the (non-)existence of stable outcomes and the way choice functions are derived from preferences.

By definition, the stability of an outcome is completely determined by agents' choice functions. In matching models where preferences are taken as primitive — such as the model we present here — the standard approach to deriving those choice functions is to

[17] We could alternatively define a block as the *full proposal* for changing the set of contracts, and replace (ii) with

ii'. There does not exist $Z \subseteq X$ such that for all $i \in N(Z \setminus Y)$, $Z_i = C_i((Z \cup Y)_i|(Z \cup Y)_{-i})$.

If we did, our stability concept would generalize *weak setwise stability* (Klaus and Walzl (2009)), rather than stability, to account for externalities. These definitions are equivalent with strategic consistency: Whenever a block (in the sense of (ii)) is successful, all agents agree about the set of contracts that will obtain after it occurs (as they must in (ii')). But with nonstrategic choice, (ii') is stronger than (ii), and so replacing (ii) with it weakens the definition of stability. We discuss this point in greater detail in Section S.2 of the Online Appendix.

[18] We generalize the usual definition of stability to accommodate externalities in a slightly different way than Pycia and Yenmez (2023) do in their two-sided setting. Under the definition they adopt, agents in a blocking coalition do not anticipate any changes to the set of contracts signed by other agents, even the other members of the blocking coalition. (That is, when evaluating a block Z of Y, the second argument of the choice function in their concept is Y_{-i}, rather than $Z_{-i} \cup Y_{-i}$.) Our stability definition instead assumes that agents in a blocking coalition account for the contracts added by the other agents in the coalition.

let them be the agents' favorite subsets of the available contracts.[19] With externalities, this usually becomes their favorite subset conditional on the enactment of whatever set of contracts they take as given for everyone else. That is,

$$\hat{C}_i(Y_i|Y_{-i}) := \arg\max_{S \subseteq Y_i} u_i(S \cup Y_{-i}) \text{ for each } Y \subseteq X. \tag{1}$$

When they make choices this way, agents behave as if each available contract will go into effect if they choose it. But for a contract to go into effect, it must *also* be chosen by *other* agents. Hence, when agents are equipped with the choice functions described in (1), they implicitly assume that *all contracts that are available will actually be chosen by the other agents they name.* Our key observation is that the standard approach does not always yield a stable outcome precisely because these assumptions may be incorrect. A familiar example illustrates this point.

Example 1 (Roommate Problem). Consider the classical roommate problem from Gale and Shapley (1962). Three friends must come to an agreement about which two of them will rent an apartment together: $I = \{1, 2, 3\}$, $X = \{x_{12}, x_{23}, x_{31}\}$, and $N(x_{ij}) = \{i, j\}$ for each $i, j \in I$. There are no externalities. Each agent prefers having any roommate to being unmatched, and cannot be part of two roommate agreements: $u_i(\{x_{ij}\}) > u_i(\emptyset) > u_i(\{x_{ij}, x_{ik}\})$ for each $i \in I$ and each $j \neq k \neq i$. Moreover, agents' preferences over roommates form a cycle: $u_1(x_{12}) > u_1(x_{31})$, $u_2(x_{23}) > u_2(x_{12})$, and $u_3(x_{31}) > u_3(x_{23})$.

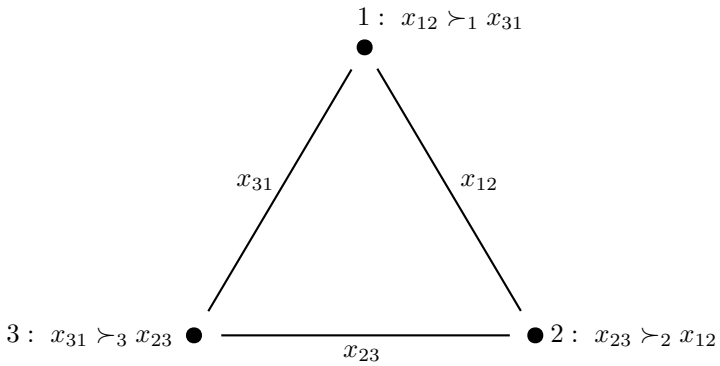

Figure 1: A visual description of the environment in Example 1.

Part 1. Suppose that preferences translate into choices in the standard way, and each agent's choice function is defined by (1). Then there is a *blocking cycle* that makes every outcome unstable: When the set of available contracts is $\{x_{12}, x_{31}\}$, agent 1 chooses x_{12}

[19] An alternative approach is to directly take choice functions as primitive, as in, e.g., Pycia and Yenmez (2023); Aygün and Sönmez (2012); or (on one side) Hatfield and Milgrom (2005).

(and rejects x_{31}), while agent 2 chooses the only agreement available to him, x_{12}. Hence, $\{x_{12}\}$ blocks $\{x_{31}\}$ — and by symmetry, $\{x_{31}\}$ blocks $\{x_{23}\}$ and $\{x_{23}\}$ blocks $\{x_{12}\}$. Since no outcome with more than one contract can be individually rational, and $\emptyset$ is blocked by any $\{x_{ij}\}$, we arrive at the standard conclusion that — with choice functions defined as in (1) — *the roommate problem has no stable outcome.*

This absence of a prediction can be attributed to implicit assumptions by the agents that happen to be mistaken. To see this, first suppose that *all three* agreements are available. Each of the three friends now has access to their favorite roommate agreement, so with the standard approach to choice described in (1), that is precisely the agreement they choose:[20]

$$\hat{C}_1(X_1|X_{-1}) = \{x_{12}\}; \qquad \hat{C}_2(X_2|X_{-2}) = \{x_{23}\}; \qquad \hat{C}_3(X_3|X_{-3}) = \{x_{31}\}. \tag{2}$$

None of these choices coincide: Every agreement is rejected by someone, and so no pair agrees to room together even though each agent would prefer rooming with anyone to remaining alone.[21] This is because agents implicitly made incorrect assumptions about each other's choices. For instance, if agent 1 had correctly anticipated the other agents' choices, he would have chosen $\{x_{31}\}$ instead. When he chose $\hat{C}_1(X_1|X_{-1}) = \{x_{12}\}$ according to (1), he implicitly assumed that each of x_{12} and x_{31} would take effect if he chose it. In particular, he was also assuming that when all agreements were available, x_{31} would be chosen by agent 3 (which turned out to be correct), and x_{12} would be chosen by agent 2 (which turned out to be incorrect).

Part 2. But what if we replaced agents' incorrect assumptions with correct beliefs about others' choices, and required those beliefs to be consistent *across* all sets of available contracts? Then the blocking cycle vanishes, and a stable outcome exists.

For instance, suppose that when all agreements are available, agent 1 believes that, as in (2), x_{31} will be chosen by agent 3, but x_{12} will be rejected by agent 2. Then his optimal choice is $C_1(X_1|X_{-1}) = \{x_{31}\}$. If agent 3 correctly believes that x_{31} will be chosen by agent 1, then since it is his favorite contract, he will optimally choose it as well: $C_3(X_3|X_{-3}) = \{x_{31}\}$. And if agent 2 correctly believes that neither of his friends will choose an agreement with him, it is optimal for him to choose nothing: $C_2(X_2|X_{-2}) = \emptyset$. Hence, the beliefs agent 1 had about the contracts that agents 2 and 3 would choose are correct.

Since they are correct, these beliefs eliminate the myopic behavior observed in Part 1 when all three contracts were available. If they are consistent with beliefs about choices

[20] Recall that there are no externalities in this environment, so the presence of x_{jk} does not change agent i's preferences over subsets of $X_i = \{x_{ij}, x_{ik}\}$.

[21] This does not mean that autarky ($\emptyset$) is a stable outcome; on the contrary, no outcome is stable. It merely means that it is the outcome chosen by nonstrategic agents when all agreements are available.

from *other* sets of contracts, they also restore the existence of a stable outcome. Recall that with all three contracts available, none of the agents believed that any of the others would choose x_{23}. If beliefs are consistent across sets of available contracts, making x_{23} unavailable should not change agents' beliefs about the *remaining* contracts $\{x_{12}, x_{31}\}$. Hence, choices should not change either: we should have $C_1(\{x_{12}, x_{31}\}|\emptyset) = C_3(\{x_{31}\}|\{x_{12}\}) = \{x_{31}\}$, and $C_2(\{x_{12}\}|\{x_{31}\}) = \emptyset$, and so $\{x_{31}\}$ blocks $\{x_{12}\}$, rather than the other way around.

These choices break the blocking cycle that ruled out the existence of a stable outcome. In fact, if we continue to construct agents' choice functions in this manner — as optimal choices given correct beliefs that are consistent across sets of available contracts — we arrive at a profile for which $\{x_{31}\}$ is the *unique* stable outcome.[22] (Other strategically consistent profiles can be found for which the stable outcome is different (e.g., $\{x_{23}\}$ or $\{x_{12}\}$); see Example 2.) Even though agents 1 and 2 would prefer the outcome to be $\{x_{12}\}$ rather than $\{x_{31}\}$, they fail to coordinate on a block because both correctly believe that the other agent would not follow through with it. This can be rationalized by a simple story: Agent 1 (correctly) believes that if he breaks his agreement with agent 3 to room with agent 2, agent 2 will then leave to form an agreement with the newly roommateless agent 3.[23] ■

In this section, we show that Example 1's conclusion holds more generally. Suppose that we account for each agents i's *beliefs* about the contracts that other agents will choose, in the form of a mapping $\mu_i : 2^X \to 2^X$ that returns the set of contracts $\mu_i(Y)$ that he believes *no other agent will reject* from an available set of contracts Y. This is a sufficient statistic for his beliefs about the choices of the other agents: all that matters for his own choice is which contracts will go into effect if he chooses them.

Our main results show that if we derive agents' choice functions given their beliefs, then a stable outcome exists in *any* matching environment, even without restrictions on the agents' preferences or the market structure, whenever beliefs are *correct* (match other agents' actual choices) and *cross-set consistent* (match beliefs at sets from which irrelevant contracts are

[22] The choices and beliefs pinned down above (those when all contracts are available and when $\{x_{12}, x_{31}\}$ are available) pin down the stable outcome as $\{x_{31}\}$. They are consistent with multiple profiles of optimal choices and correct beliefs at *other* sets of available contracts, and while each of these profiles has the same stable outcome, they may lead to different comparative statics (e.g., if $\{x_{31}\}$ were removed). One such profile is given by

$$\begin{array}{lll} C_i(\emptyset|\emptyset) = \emptyset; & C_i(\{x_{ij}\}|\emptyset) = \{x_{ij}\}; & C_i(\emptyset|\{x_{jk}\}) = \emptyset, \text{ for each } i \neq j \neq k; \\ C_1(\{x_{31}\}|\{x_{23}\}) = \{x_{31}\}; & C_2(\{x_{23}\}|\{x_{31}\}) = \emptyset; & C_3(\{x_{31}, x_{23}\}|\emptyset) = \{x_{31}\}; \\ C_1(\{x_{12}\}|\{x_{23}\}) = \emptyset; & C_2(\{x_{12}, x_{23}\}|\emptyset) = \{x_{23}\}; & C_3(\{x_{23}\}|\{x_{12}\}) = \{x_{23}\}. \end{array}$$

Intuitively, we can think of the players' beliefs about one another's choices in each of these profiles as reflecting the players' relative bargaining power. We explore this connection in greater detail in Section 5.

[23] We explore the consequences of requiring profiles to be rationalized by such *forward induction* reasoning in the Online Appendix.

removed). We call such a profile of choice functions and beliefs *strategically consistent*.

Definition (Strategic Consistency and Nonstrategic Choice). Given agents' payoffs $\{u_i : 2^X \to \mathbb{R}\}_{i \in I}$,

- A profile of choice functions $\{C_i : 2^{X_i} \times 2^{X_{-i}} \to 2^{X_i}\}_{i \in I}$ and beliefs $\{\mu_i : 2^X \to 2^X\}_{i \in I}$ is *strategically consistent* if for each $i \in I$,

  i. μ_i is *correct* given $\{C_j\}_{j \neq i}$: For each $Y \subseteq X$, it holds that $\mu_i(Y) = C_{-i}(Y) := \bigcap_{j \neq i}(C_j(Y_j|Y_{-j}) \cup Y_{-j})$.[24]

  ii. C_i is *optimal* given μ_i: For each $Y \subseteq X$, it holds that $C_i(Y_i|Y_{-i}) = \arg\max_S u_i(S \cup \mu_i(Y)_{-i})$ s.t. $S \subseteq \mu_i(Y)_i$.

  iii. μ_i is *cross-set consistent* given $\{C_i\}_{i \in I}$: For each $Y, Z \subseteq X$, if $Y \supseteq Z \supseteq C_j(Y_j|Y_{-j})$ for all $j \in I$, then $\mu_i(Z) = \mu_i(Y)$.

- Each agent i's *nonstrategic choice function* $\hat{C}_i$ is defined by (1).

Strategic consistency is motivated by two assumptions about agents' epistemic sophistication. First, when faced with any set of contracts that might be proposed, they are able to form correct beliefs about which contracts the other agents will choose. Second, when contracts that are not chosen by *anyone* are removed, agents do not believe that others would change their behavior (cross-set consistency). (E.g., in Example 1, since no agent chose x_{23} when all contracts were available, we required that none of them would change their choices when we made x_{23} unavailable.) That is, when agents negotiate, those negotiations are independent of irrelevant alternatives. This criterion has bite because strategic consistency requires agents to form beliefs at *each set of available contracts, not just those that are involved in blocks of a potentially stable outcome.*[25]

When agents instead assume that each contract they choose will go into effect, we call the resulting choice functions — those derived from preferences using the standard approach — *nonstrategic*.

[24] The reason that the set of contracts $C_{-i}(Y)$ not rejected by the other agents is defined this way, rather than as the intersection $\bigcap_{j \neq i} C_j(Y_j|Y_{-j})$ of the contracts *chosen* by each agent $j \neq i$, is because each $C_j(Y_j|Y_{-j})$ is a subset of Y_j, and so any contract in Y that did not name *every* agent $j \neq i$ would not be in the intersection $\bigcap_{j \neq i} C_j(Y_j|Y_{-j})$. Instead, since not every agent has an opportunity to choose every contract, we have to intersect contracts *not rejected* by $j \neq i$, i.e., $C_j(Y_j|Y_{-j}) \cup Y_{-j}$.

[25] Observe that in Example 1, $\{x_{31}\}$ can only be blocked by $\{x_{12}\}$ or $\{x_{23}\}$ *alone* since agent 2 cannot sign both agreements. (The important part here is that agent 2 *would never* choose both agreements; the interpretation in the roommate example just so happens to be that doing so is infeasible.) But cross-set consistency ruled out blocking cycles — thus ensuring a stable outcome — precisely because agents had correct beliefs about each other's choices when all of the contracts were available *together*.

Remark. The fixed point between choices and beliefs described by (i) and (ii) is reminiscent of Nash equilibrium. This analogy can be made precise: at every set of available contracts Y, agents' choice functions must return a Nash equilibrium of a game in which each agent announces a subset $S_i \subseteq Y_i$, and a contract $x \in Y$ goes into effect if it is announced by every agent $i \in N(x)$ that it names.[26,27] (We formalize this connection in Section S.5 of the Online Appendix.) This does not mean that the combination of stability and strategic consistency is just a noncooperative equilibrium concept cast in matching-theoretic language. Instead, it requires robustness to *joint* deviations given *equilibrium* behavior by individuals after each potential deviation that might be proposed — i.e., both "on" and "off" the "equilibrium path".

3.1 Stable Outcomes

Our first main result shows that each strategically consistent profile of choice functions and beliefs pins down a stable outcome. Because agents make correct assumptions about each other's behavior, none of the conditions used to show that stable outcomes exist with nonstrategic choice — e.g., substitutable preferences, no (or well-behaved) externalities, acyclic or two-sided market structure — are necessary to ensure that stable outcomes exist. We say an outcome $Y \subseteq X$ is *stable for* a profile $\{C_i, \mu_i\}_{i \in I}$ if it is stable given choice functions $\{C_i\}_{i \in I}$.

Theorem 1 (Strategic Consistency and Stability). *For each strategically consistent profile of choice functions and beliefs $\{C_i, \mu_i\}_{i \in I}$, there is a unique outcome that is stable for that profile.*

There is a straightforward reason that strategically consistent choice ensures a stable outcome in settings where nonstrategic choice does not. Essentially, nonstrategic choice can permit agents to coordinate on *too many* blocks for a stable outcome to exist. Theorem 1 shows that — as illustrated in Example 1 — when it does, these blocks either are not internally consistent (in the sense that some agent would not choose to participate in them

[26] This *contract-announcement game* generalizes the link-announcement game discussed in, e.g., Myerson (1991) and Jackson (2010).

[27] Note that for the same reason that weakly dominated strategies may be played in Nash equilibrium, agents' choice functions may return "weakly dominated" choices at some available sets of contracts — e.g., agent 2 may choose $\emptyset$ instead of $\{x_{12}\}$ in the roommate problem when $\{x_{12}, x_{31}\}$ are available — when the choice functions are part of a strategically consistent profile. In some environments (such as the roommate problem), this happens at some set of available contracts in *any* strategically consistent profile. Intuitively, while we can find an equilibrium of the contract-announcement game at any *given* set of available contracts in which no agent plays a weakly dominated strategy, these equilibria need not be consistent with each other across *different* sets of available contracts (in the sense of cross-set consistency); see Section 5 for a discussion.

if they had correct beliefs about the other agents' choices) or are not consistent with each other (in the sense of cross-set consistency). That is, the assumption that agents can *always* successfully coordinate on a block necessarily introduces inconsistencies in their choice behavior. Resolving these inconsistencies with a strategically consistent profile then pins down a unique stable outcome.

All three parts of strategic consistency play a role in this result. First, correctness and optimality ensure that agents have common beliefs that match each other's choices. Formally:

Lemma 1. *Suppose choice functions $\{C_i\}_{i\in I}$ are optimal given beliefs $\{\mu_i\}_{i\in I}$, and beliefs $\{\mu_i\}_{i\in I}$ are correct given choice functions $\{C_i\}_{i\in I}$. Then for each $i, j \in I$ and $Y \subseteq X$,*

i. Agents' beliefs must coincide: $\mu_i(Y) = \mu_j(Y) =: \mu(Y)$.

ii. Agents' choices match common beliefs: $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$.

Intuitively, when choice is optimal given correct beliefs, no agent ever chooses a contract from a set of available contracts that another agent rejects from that set; i.e., unlike in Part 1 of Example 1, contracts must either be rejected by *everyone* they name or rejected by *no one*. Consequently, since all agents' beliefs are correct, they must (i) coincide and (ii) match the choices of *each* individual agent, not just the set of contracts that *none* of them reject (as in the definition of correct beliefs).

Second, since choices match a common belief at *each* set of available contracts, consistency of those beliefs *across* sets of available contracts rules out the kind of blocking cycles that can lead to nonexistence when choice is nonstrategic (as in Part 1 of Example 1). In particular, it ensures that whatever set $\mu(X)$ agents believe others will choose from the set of *all* contracts X, they also believe others will choose $\mu(X)$ from any set $Y \supseteq \mu(X)$ that contains it.

This guarantees that each agent chooses precisely the contracts in $\mu(X)$ that name them when $Y = \mu(X)$ (individual rationality), and chooses no new contracts Z that might be available alongside it when $Y = \mu(X) \cup Z$ (unblocked). Hence, $\mu(X)$ is stable for the profile $\{C_i, \mu_i\}_{i\in I}$. It also guarantees that $\mu(X)$ is the *unique* stable outcome for that profile: any $S \neq \mu(X)$ either isn't individually rational (if $S \supset \mu(X)$) or is blocked by $\mu(X) \setminus S$ (otherwise).

Thus, Theorem 1 is constructive: once we have found a strategically consistent profile, it is straightforward to find the unique outcome that is stable for that profile, since it coincides with any agent's beliefs $\mu_i(X)$ when all contracts are available. Consequently, with strategic consistency, instead of finding stable outcomes, we can direct our efforts toward finding and characterizing profiles of choice functions and beliefs.

Corollary 1 (Stability and Beliefs). *Given a strategically consistent profile* $\{C_i, \mu_i\}_{i \in I}$*,* Y *is the unique stable outcome for that profile of choice functions and beliefs if and only if* $\mu_i(X) = Y$ *for each* $i \in I$*.*

The process by which these outcomes are formed is simple. Each agent forms correct and consistent beliefs about the way that other agents deviate both noncooperatively (i.e., by dropping contracts unilaterally) and cooperatively (i.e., by responding to proposed blocks). Then, they each agree to the contracts that are part of the unique set that is robust to these deviations.

3.2 Strategically Consistent Profiles

Theorem 1 shows that each strategically consistent profile of choice functions and beliefs pins down a unique stable outcome. However, the existence of these profiles is not immediate: Strategically consistent profiles are equilibrium objects, in the sense that given a set of contracts, each agent makes optimal choices, given the choices of the others.

Our second main result shows that strategically consistent profiles always exist in any matching environment.

Theorem 2 (Strategically Consistent Profiles: Existence). *Strategically consistent profiles exist.*

Theorem 2 establishes the existence of fixed points in *choice functions* (i.e., strategically consistent profiles) rather than fixed points in *outcomes* (e.g., the outcomes of a deferred acceptance algorithm). To explain it, we describe the algorithm that we introduce to construct strategically consistent profiles of choice functions and beliefs.

We start by considering the outcomes Y that are *nonstrategically individually rational*: $\hat{C}_i(Y_i|Y_{-i}) = Y_i$ for each $i \in I$. These outcomes play an important role in the construction of strategically consistent profiles: they are precisely the sets of contracts that agents can believe others will choose from an available set of contracts.[28] This fact facilitates a converse to Lemma 1 that powers our construction algorithm.

Lemma 2 (Converse of Lemma 1). *Suppose that* $\{C_i, \mu_i\}_{i \in I}$ *is a profile of choice functions and beliefs such that beliefs are common across agents, and choices match beliefs: For each* $i, j \in I$ *and* $Y \subseteq X$*,* (i) $\mu_i(Y) = \mu_j(Y) := \mu(Y)$*, and* (ii) $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$*. Then*

[28] Intuitively, when choice functions are optimal given correct beliefs, those beliefs must be common across agents (Lemma 1). Then at any set of available contracts Z, no one can have an incentive to reject contracts that are part of the common belief $\mu(Z)$, given that the other agents choose precisely the contracts in $\mu(Z)$. In other words, $\mu(Z)$ must be nonstrategically individually rational.

(a) The beliefs $\{\mu_i\}_{i\in I}$ are correct given the choice functions $\{C_i\}_{i\in I}$.

(b) The choice functions $\{C_i\}_{i\in I}$ are optimal given the beliefs $\{\mu_i\}_{i\in I}$ if and only if for each $Y \subseteq X$, $\mu(Y)$ is nonstrategically individually rational.

Our algorithm is initialized by picking some strict total order $\succ$ on the collection of nonstrategically individually rational outcomes. Then, at each each set of available contracts, have each agent choose the contracts in the highest-ranked outcome available, and correctly believe that the other agents will do the same:

$$\underbrace{\mu_i(Y) = \mu(Y) =}_{\text{beliefs are common}} \underbrace{\max_{\succ}\{Y' | Y' \subseteq Y\}}_{\substack{\succ\text{-highest nonstrategically} \\ \text{IR outcome available}}}, \qquad \underbrace{C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i.}_{\text{choices match beliefs}} \tag{3}$$

Intuitively, the order $\succ$ used in this algorithm captures the agents' common assumptions about which outcomes will result from any joint deviation that might be proposed. Because this order pins down beliefs at *every* set of available contracts, these beliefs are cross-set consistent. Since these beliefs are common across agents and match choices, the algorithm always generates a strategically consistent profile of choice functions and beliefs (Lemma 2).

3.3 Predictions

Theorems 1 and 2 show that strategic consistency and stability always allow an analyst to make predictions about outcomes in any matching environment. However, these predictions are not necessarily unique: Even though there is a unique stable outcome for any particular profile of choice functions and beliefs (Theorem 1), there may be multiple strategically consistent profiles, and different outcomes may be stable for different profiles. Intuitively, strategically consistent profiles resolve inconsistencies in agents' choices, and there are generally multiple ways in which those inconsistencies can be resolved. If we remain agnostic about which strategically consistent profile will describe the agents' behavior, then, what is the set of outcomes we can predict?[29]

Proposition 1 (Predictions of Strategic Consistency). *There is a strategically consistent profile for which the outcome $Y \subseteq X$ is stable if and only if Y is nonstrategically individually rational.*

[29]Recall from Corollary 1 that for any given strategically consistent profile, the unique outcome stable for that profile is pinned down by the agents' common belief at X, $\mu(X)$. Thus, that outcome must be nonstrategically individually rational (Lemma 2). In fact, nonstrategic individual rationality is also sufficient: Proposition 1 shows that we can use our algorithm to construct a profile for which such an outcome is stable merely by picking an order $\succ$ that ranks it highest.

In particular, there is always a strategically consistent profile for which the autarky outcome $\emptyset$ is stable. (See Example 2.)

In a sense, Proposition 1 is unsatisfying. It seems clear that a coherent theory of agreement formation should rule out outcomes where an agent can benefit by dropping agreements they are involved in — i.e., those outcomes that are not (nonstrategically) individually rational. The purpose of the stability concept is to *also* rule out outcomes that are not robust to joint deviations. Proposition 1 shows that if, after any proposed deviation, agents take the choices of their counterparties into account by forming correct beliefs about them, then unless we focus on a specific profile or profiles of beliefs, *robustness to joint deviations provides no additional predictive power.*

But in another sense, Proposition 1 doesn't describe a problem with strategic consistency so much as it describes a consequence of remaining agnostic about agents' beliefs. Thus, in the remainder of the paper, we explore ways to select among profiles of beliefs instead.

4 Refinements

The appropriate way to narrow the set of strategically consistent profiles under consideration can be informed by context.

For instance, suppose we observe an outcome in the data, and want to make a prediction about the outcome in a counterfactual scenario (e.g., after a merger or the imposition of regulation). A common approach to this problem is to use the observed outcome to estimate a set of parameters (e.g., Nash bargaining weights (Crawford and Yurukoglu, 2012; Grennan, 2013; Ho and Lee, 2017)), and then use those parameters to predict the counterfactual outcome. As Rostek and Yoder (2025) shows, one can pin down a strategically consistent profile (and thus an outcome) in the counterfactual environment in an analogous way, using a profile of beliefs recovered from the outcome that we observe in the data.

Alternatively, suppose that we want to use stability and strategic consistency as a tool to rule out outcomes without inferring a specific profile of choice functions and beliefs from data. As Example 2 illustrates, some profiles of beliefs may be more intuitively plausible than others. This suggests that we should develop criteria for ruling out profiles with less plausible beliefs — and thus outcomes that require such beliefs to be stable. This is what we do in this section.

Example 2 (Roommate Problem Revisited). Consider the roommate problem from Example 1 once more. There, we found a strategically consistent profile for which $\{x_{31}\}$ was

stable: $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$ and $\mu_i(Y) = \mu(Y)$, for each $i \in \{1,2,3\}$ and

$$\mu(\emptyset) = \emptyset; \qquad \mu(\{x_{ij}\}) = \{x_{ij}\} \text{ for each } i,j;$$
$$\mu(\{x_{31}, x_{23}\}) = \mu(\{x_{12}, x_{31}\}) = \{x_{31}\}; \qquad \mu(\{x_{12}, x_{23}\}) = \{x_{23}\}.$$

By symmetry, profiles also exist for which $\{x_{12}\}$ and $\{x_{23}\}$ are stable. In these profiles, people find roommates, in line with the experience of most undergraduate students.

But there is also a strategically consistent profile for which the "autarky" outcome $\emptyset$ is stable: $\{C_i^0, \mu_i^0\}_{i \in I}$, where $C_i^0(Y_i|Y_{-i}) = \mu_i^0(Y) = \emptyset$ for all $Y \subseteq X$ and $i \in \{1,2,3\}$. Here, agents always reject each roommate agreement x_{ij} that might be proposed, because they correctly believe that their prospective roommate will also reject it.

This seems less intuitively plausible, for multiple reasons.

- ♦ In $\{C_i^0, \mu_i^0\}_{i \in I}$, agents fail to coordinate on choosing a roommate agreement x_{ij} when one is made available alongside $\emptyset$, even though the resulting outcome $\{x_{ij}\}$ would be nonstrategically individually rational and a Pareto improvement. This is similar to the kind of coordination failure that motivates *setwise stability* in two-sided many-to-many matching (Klaus and Walzl, 2009).[30] Thus, unlike $\{C_i, \mu_i\}_{i \in I}$, $\{C_i^0, \mu_i^0\}_{i \in I}$ resolves inconsistencies in nonstrategic choice by encoding coordination failures in which agents choose a Pareto-dominated set of contracts, even though they could choose a Pareto-improving outcome and still satisfy strategic consistency.[31] In contrast, the coordination failure encoded in $\{C_i, \mu_i\}_{i \in I}$ — agents 1 and 2 fail to block $\{x_{31}\}$ — can be rationalized by agent 1 foreseeing that because it would lower agent 3's utility, such a block makes it profitable for both agents 2 and 3 to agree to $\{x_{23}\}$, leaving him unmatched.

- ★ In $\{C_i^0, \mu_i^0\}_{i \in I}$, agent i rejects any roommate agreement x_{ij} that might be proposed by an agent $j \neq i$ as a block of $\emptyset$, even though agent i should view such a block as *credible*: agent j could not benefit by proposing it and then not choosing it, and given the profile $\{C_i^0, \mu_i^0\}_{i \in I}$, no agent could benefit by proposing it as an intermediate step toward an anticipated further deviation. Thus, unlike $\{C_i, \mu_i\}_{i \in I}$, $\{C_i^0, \mu_i^0\}_{i \in I}$ is inconsistent with forward induction reasoning by the agents.[32] ■

[30] Unlike the usual, nonstrategic approach to stability, setwise stability allows agents to coordinate on an individually rational block even when it is not their favorite subset available from among the blocking contracts and the existing contracts.

[31] No such coordination failures exist in $\{C_i, \mu_i\}_{i \in I}$ (or the symmetric profiles for which $\{x_{12}\}$ and $\{x_{23}\}$ are stable): At any set of available contracts, there is no outcome that represents a Pareto improvement upon the outcome that the agents believe the others will choose.

[32] $\{C_i, \mu_i\}_{i \in I}$ is consistent with forward induction reasoning: Even though agents 1 and 2 could benefit

In Section 4.1, we extend the reasoning in ♦ to a *Pareto optimality* refinement that can be applied more generally. In Section 4.2, we develop two refinements — *forward induction* and *weak forward induction* — based on the reasoning described in ★.

4.1 Pareto Optimality

In this section, we first introduce a criterion on strategically consistent profiles ensuring that agents' beliefs select Pareto-undominated outcomes whenever possible. We then give a "welfare theorem" characterizing the stable outcomes predicted by these profiles. As it turns out, this also makes precise the connection between strategically consistent profiles and Nash bargaining weights.

Recall that when choices are optimal, correct beliefs always select nonstrategically individually rational subsets from each set of available contracts. Our main efficiency criterion requires that they never select one of these subsets when it is Pareto-dominated by another. That is, it focuses our attention on profiles that resolve inconsistencies in nonstrategic choice by ruling out successful coordination *only* when that coordination would not be Pareto-improving.

Definition (Pareto Optimality). We say that a strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$ satisfies *Pareto optimality* if for any nonstrategically individually rational $Y, Z \subseteq X$ such that $u_i(Y) \geq u_i(Z)$ for all $i \in I$ and $u_i(Y) > u_i(Z)$ for some $i \in I$, we have $\mu_i(Y \cup Z) \neq Z$ for each $i \in I$.

The Pareto optimality refinement is weak. Instead of assuming that agents will succeed in coordinating on a nonstrategically individually rational block whenever the *subset* of agents that sign new contracts prefer it (as we would if we applied, e.g., setwise stability), it merely assumes that they will succeed in coordinating on such a block when *all* agents prefer it.

Theorem 3 shows that strategically consistent profiles that satisfy Pareto optimality are easy to find, simply by using our algorithm (3) with an order $\succ$ that is structured so that the agents' common beliefs solve a social planner's problem. Formally, we say that a strict total order $\succ^\phi$ on the nonstrategically individually rational outcomes $\mathcal{M}$ is *induced by* $\phi : \mathbb{R}^I_+ \to \mathbb{R}$ if $\phi((u_i(Y))_{i \in I}) > \phi((u_i(Z))_{i \in I})$ implies $Y \succ^\phi Z$. As Lemma 9 in the Appendix shows, each increasing ϕ induces an order $\succ^\phi$. We can interpret ϕ (and the $\succ^\phi$ it induces) as describing the way that agents base their beliefs about other agents' choices on the payoffs all agents will receive from those choices.

directly from a block of $\{x_{31}\}$ by $\{x_{12}\}$, that block is not credible in that profile, since agent 2 can benefit by proposing it as an intermediate step to forming an agreement with agent 3.

Theorem 3 (Pareto-Optimal Profiles). *Let $\phi : \mathbb{R}^I_+ \to \mathbb{R}$ be a strictly increasing function.*

i. For any strict total order $\succ^\phi$ induced by ϕ, the profile $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$ constructed from $\succ^\phi$ using the algorithm (3) is strategically consistent and satisfies Pareto optimality.

ii. The common belief μ^ϕ in $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$ is the solution to a social planner's problem: For all $i \in I$ and $Z \subseteq X$,

$$\mu_i^\phi(Z) \in \arg\max_{S\subseteq Z} \phi((u_i(S))_{i\in I}) \text{ s.t. } \hat{C}_j(S_j|S_{-j}) = S_j \ \forall j \in I. \tag{4}$$

The intuition for Theorem 3 is straightforward. Since ϕ is strictly increasing, $\succ^\phi$ ranks Y ahead of Z whenever Y is a Pareto improvement on Z. Hence, the beliefs constructed by the algorithm never select a Pareto-inferior outcome when a Pareto-superior one is available (i). In fact, the construction of $\succ^\phi$ ensures that they solve the social planner's problem (4) (and do so uniquely if there are no ties) (ii).

Theorem 3 shows that beliefs are part of a strategically consistent profile if, at any set of available agreements, they maximize a social welfare function subject to the constraint that no agent can profit by vetoing contracts (and ties are broken in a consistent manner). This provides one direction of a "welfare theorem" for strategic consistency (Theorem 4): outcomes that are stable for some strategically consistent profile satisfying Pareto optimality are those on a *constrained* Pareto frontier (where the constraint is individual rationality).[33]

Theorem 4 (Welfare Theorem for Strategic Consistency). *There is a strategically consistent profile satisfying Pareto optimality for which $Y \subseteq X$ is stable if and only if Y is Pareto efficient among the nonstrategically individually rational outcomes.*

Theorem 4 parallels welfare theorem-like results in the matching literature with *transferable utility* (e.g., with substitutability, Hatfield et al. (2013, Theorems 2-6); with complementarity, Rostek and Yoder (2020, Proposition 3 and Theorem 2)). Rather than decentralizing an efficient outcome through the use of *competitive equilibrium prices*, Theorem 4 shows that such outcomes can be decentralized by a correct and consistent profile of (common)

[33] Pareto efficient outcomes may fail to be nonstrategically individually rational. In fact, it may be that there is *no* efficient outcome that is nonstrategically individually rational. For instance, if $I = \{1, 2\}$, $X = \{x, y\}$, $N(x) = N(y) = I$, and the agents' preferences are given by

$$\{x\} \succ_1 \{x, y\} \succ_1 \emptyset \succ_1 \{y\}$$
$$\{y\} \succ_2 \{x, y\} \succ_2 \emptyset \succ_2 \{x\},$$

then $\{x, y\}$, $\{x\}$, and $\{y\}$ are all Pareto efficient, but only $\emptyset$ is nonstrategically individually rational.

beliefs.[34] This allows it to apply even to nontransferable utility settings, and without the conditions on preferences or market structure that are, in general, necessary for competitive equilibrium prices to exist.[35] Example 3 illustrates in a many-to-one matching model with both complementarity and substitutability, where the standard, nonstrategic approach to choice does not yield a stable outcome.

Example 3 (Labor Markets with Complementarities). Consider a labor market with two workers, Alice and Bob, and two firms, 1 and 2: $I = \{a, b, 1, 2\}$. Employment contracts are standardized, i.e., each is completely characterized by the worker-firm pair it involves: $X = \{x_{a1}, x_{a2}, x_{b1}, x_{b2}\}$ and $N(x_{ij}) = \{i, j\}$ for each $i \in \{a, b\}$ and $j \in \{1, 2\}$.

Workers can only work for one firm ($u_i(\{x_{i1}, x_{i2}\}) < u_i(\emptyset)$ for each $i \in \{a, b\}$), and there are no externalities. Both workers prefer any employment to unemployment, but Alice prefers firm 1, while Bob prefers firm 2: $u_a(\{x_{a1}\}) > u_a(\{x_{a2}\})$ and $u_b(\{x_{b2}\}) > u_b(\{x_{b1}\})$. Firm 2 wants to hire one worker ($u_2(\{x_{a2}, x_{b2}\}) < u_2(\emptyset)$), and would prefer it to be Alice: $u_2(\{x_{a2}\}) > u_2(\{x_{b2}\})$. Firm 1 could hire both workers, but is only willing to hire Alice if it also hires Bob: $u_1(\{x_{a1}, x_{b1}\}) > u_1(\{x_{b1}\}) > u_1(\emptyset) > u_1(\{x_{a1}\})$.

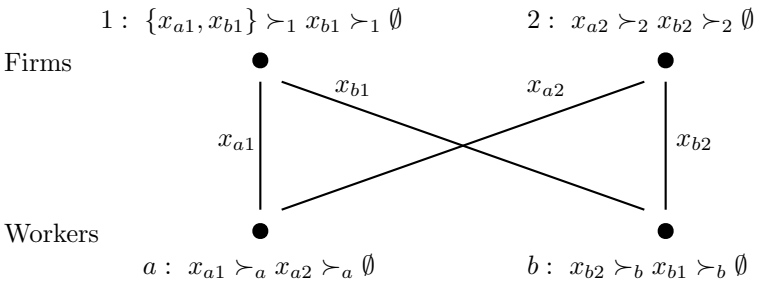

Figure 2: A visual description of the environment in Example 3.

Like many matching environments with both complementarity (between x_{a1} and x_{b1}) and substitutability (among other contracts), this example has no stable outcome when choice is nonstrategic.[36] But when choice is strategically consistent, and agents can overcome coordination failures that result in Pareto-dominated outcomes, Theorem 4 shows that three outcomes can be stable: $\{x_{b2}\}$, $\{x_{a2}, x_{b1}\}$, and $\{x_{a1}, x_{b1}\}$. Each is nonstrategically

[34] In several matching papers with transferable utility (e.g., Hatfield et al. (2013)), competitive equilibrium is used as a tool to link stability with efficiency, and thereby show that stable outcomes exist. While matching markets do not necessarily feature publicly posted prices, competitive equilibrium remains a useful benchmark in these environments. Welfare theorems show that agents act *as if* there were prices that, like the *common* beliefs in our paper, they had common certainty about.

[35] In particular, the "if" part does not follow immediately from the separating hyperplane theorem and Theorem 3, since the utility possibility set is finite, rather than convex.

[36] With nonstrategic choice, each individually rational outcome is blocked: $\emptyset$ is blocked, e.g., by x_{a2}; $\{x_{a2}\}$ is blocked by, e.g., $\{x_{b1}\}$; $\{x_{a2}, x_{b1}\}$ is blocked by $\{x_{a1}, x_{b1}\}$; $\{x_{a1}, x_{b1}\}$ is blocked by $\{x_{b2}\}$; $\{x_{b2}\}$ is blocked by $\{x_{a2}\}$; and $\{x_{a2}\}$ is blocked by $\{x_{a1}, x_{b1}\}$.

individually rational, and (unlike $\emptyset$, $\{x_{a2}\}$, and $\{x_{b1}\}$) is not Pareto-dominated by another nonstrategically individually rational outcome.[37]

We describe one such profile, for which $\{x_{a2}, x_{b1}\}$ is stable. To avoid assigning cardinal values to agents' payoffs, we start with a strict total order that never ranks an outcome below another that it Pareto dominates: e.g.,

$$\{x_{a2}, x_{b1}\} \succ \{x_{b2}\} \succ \{x_{a1}, x_{b1}\} \succ \{x_{b1}\} \succ \{x_{a2}\} \succ \emptyset. \tag{5}$$

Given this order, the algorithm in (3) constructs a strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$ satisfying Pareto optimality and forward induction for which $\{x_{a2}, x_{b1}\}$ is stable.[38] ■

Pareto Optimality and Bargaining Weights

Using Theorem 3 to construct a profile $\{C_i^\phi, \mu_i^\phi\}_{i \in I}$ requires us to first pick a social welfare function ϕ. We may wish to do so in a way that ensures that the order $\succ^\phi$ is not sensitive to specifics of the agents' utility functions that do not affect their incentives. In particular, we might desire the solution to (4) to be invariant under rescaling of the agents' utility functions. As is well known, this pins down the social welfare function in Theorem 3 as the familiar asymmetric Nash product. Formally, we say that $\phi : \mathbb{R}_+^I \to \mathbb{R}$ is *scale invariant* if for any $a, x, y \in \mathbb{R}_+^I$, it holds that $\phi(x) > \phi(y) \Leftrightarrow \phi((a_i x_i)_{i \in I}) > \phi((a_i y_i)_{i \in I})$.

Lemma 3. *If $\phi : \mathbb{R}_+^I \to \mathbb{R}$ is continuous, strictly increasing, and scale invariant, then there is some $\alpha \in \Delta(I)$ such that $\phi(x) \geq \phi(y) \Leftrightarrow \prod_{i \in I} x_i^{\alpha_i} \geq \prod_{i \in I} y_i^{\alpha_i}$.*

We emphasize that even when a strategically consistent profile is pinned down by maximizing a Nash product, the interpretation is *not* that the outcome is determined through a multilateral Nash bargaining among all agents over all contracts. Instead, the outcome is determined by the absence of deviations by *groups* of agents to form new contracts with one another, given beliefs that are refined by Pareto optimality.

By varying the social welfare function ϕ — e.g., by changing the bargaining weights α in a Nash product — we can construct a profile that is relatively more favorable to some

[37] In fact, since this setting has no externalities, Lemma 4 shows that each can be decentralized by beliefs that not only satisfy the Pareto optimality refinement, but also the forward induction refinements that we introduce in Section 4.2.

[38] Specifically, this profile $\{C_i, \mu_i\}_{i \in I}$ is given by $C_i(Y_i | Y_{-i}) = \mu(Y) \cup X_i$ and $\mu_i(Y) = \mu(Y)$, for

$$\begin{array}{cc} \mu(\{x_{a1}, x_{a2}\}) = \mu(\{x_{a2}\}) = \{x_{a2}\}; & \mu(Y) = \{x_{a2}, x_{b1}\} \text{ if } \{x_{a2}, x_{b1}\} \subseteq Y; \\ \mu(\{x_{b1}\}) = \{x_{b1}\}; & \mu(\emptyset) = \mu(\{x_{a1}\}) = \emptyset; \\ \mu(\{x_{a1}, x_{b1}\}) = \{x_{a1}, x_{b1}\}; & \mu(\{x_{a1}, x_{b1}, x_{b2}\}) = \mu(\{x_{a1}, x_{a2}, x_{b2}\}) = \mu(\{x_{b1}, x_{b2}\}) \\ & = \mu(\{x_{a2}, x_{b2}\}) = \mu(\{x_{a1}, x_{b2}\}) = \mu(\{x_{b2}\}) = \{x_{b2}\}. \end{array}$$

agents, and less favorable to others. This highlights the fact that strategically consistent profiles play the same role in matching-theoretic stability that bargaining weights do in other models of bargaining (e.g., Nash-in-Nash (Collard-Wexler et al. (2019)) bargaining): There are several possible profiles of weights/choice functions and beliefs, and given any such profile, the bargaining solution/stability pins down a unique prediction. Theorem 3 and Lemma 3 formalize this connection in the case of profiles that satisfy Pareto optimality; in Section S.1 of the Online Appendix, we illustrate the connection in the context of Example 3.

4.2 Forward Induction

Even though they are correct and cross-set consistent, beliefs may make predictions about others' responses to proposed cooperative deviations that are not justified once the credibility of those deviations is taken into account. This section refines the behavior captured by the set of strategically consistent profiles to rule out such beliefs using a notion of credibility based on forward induction: a blocking proposal is credible if no one has a reason to make the proposal without intending to follow through on it.[39] Moreover, the forward induction reasoning that it captures is *farsighted*: If an agent believes that a deviation is credible, it must not enlarge the set of outcomes that can be reached through a sequence of further deviations that are consistent with the agents' beliefs. When we rule out profiles that are not robust to such credible deviations, strategic consistency allows stability to make intuitive predictions in settings where nonstrategic choice does not yield a stable outcome — and in the network formation context, those where no pairwise stable outcome exists (e.g., the formation of a trading network in Jackson and Watts (2002) — see Example 4.).

Formally, given a profile of choice functions and beliefs $\{C_i, \mu_i\}_{i \in I}$, we say that the sequence of outcomes $\{Z^n\}_{n=0}^N$ is a *farsighted chain from Z to Z'* if $Z^0 = Z$ and $Z^N = Z'$, and for each n and each $i \in I$, $\mu_i(Z^n \cup Z^{n+1}) = Z^{n+1}$. In words, a farsighted path is a sequence of outcomes where each agent believes the $n+1$st outcome will result from a block of the nth.

Definition (Credible Blocks and Forward Induction). Given a strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$, we say that Z is a *credible blocking proposal* for Y if

i. (Nonstrategic Individual Rationality) Z is nonstrategically individually rational;

[39] In other words, a credible blocking proposal is one that no agent would choose to propose *unless their intentions in doing so, if fully understood by others, would prompt a response that justifies that deviation.* This is analogous to the kind of credible deviation discussed in, e.g., Kohlberg and Mertens (1986): one that an agent would not choose to make *unless their intentions in doing so, if fully understood by others, would prompt a response that justifies that deviation.*

ii. (Myopic Credibility) even if agents myopically believe that others will leave their existing contracts intact, each will agree to the new contracts in Z, and reject their old contracts in $Y \setminus Z$: $Z_i = \hat{C}_i((Y \cup Z)_i | (Y \cup Z)_{-i})$ for each $i \in I$;[40]

iii. (Farsighted Credibility) adopting the new set of contracts cannot lead to new deviations *farsightedly*: If there is a farsighted chain from Z to Z', there is a farsighted chain from Y to Z'.

We say that $\{C_i, \mu_i\}_{i \in I}$ satisfies *forward induction* if for every nonstrategically individually rational outcome Y, and every Z that is a credible blocking proposal for Y, we have $\mu_i(Y \cup Z) = Z$ for each $i \in I$. If this property holds whenever $Y \subset Z$, we say that $\{C_i, \mu_i\}_{i \in I}$ satisfies *weak forward induction*.

Blocking proposals are credible if no agent has a reason to make them unless they intend to enact the newly proposed set of contracts. In particular, they cannot benefit *directly* by misleading others about their intentions, and either keeping old contracts or rejecting the new ones that are part of the blocking proposal (myopic credibility); and they cannot benefit *indirectly* by subsequently deviating unilaterally (nonstrategic individual rationality) or jointly (farsighted credibility).

Weak forward induction requires each agent to believe that the members of a blocking coalition will go along with credible proposals to *add* contracts. Forward induction requires agents to also believe in credible proposals to *change* the set of contracts, i.e., add some contracts while deleting others.[41]

In general, our forward induction refinements are neither stronger nor weaker than Pareto optimality. But Pareto-optimal profiles must satisfy forward induction when externalities are absent, since in that case, any myopically credible blocking proposal is also a Pareto improvement.

[40] If we weakened myopic credibility to only require that all those involved in signing new contracts in Z or dropping old contracts in $Y \setminus Z$ will choose precisely the contracts in Z (i.e., for each $i \in N(Z \setminus Y)$, $Z_i = \hat{C}_i((Y \cup Z)_i | (Y \cup Z)_{-i})$, and for each $x \in Y \setminus Z$, there exists $i \in N(x)$ with $Z_i = \hat{C}_i((Y \cup Z)_i | (Y \cup Z)_{-i})$), all of our results would go through unchanged, with the exception of Lemma 4: For Theorem 5, observe that $Z \supset Y$ is nonstrategically individually rational iff it is a myopically credible block of Y under the stronger definition given here, while for Theorem 6, the key fact about myopic credibility is that if Z is a myopically credible block of Y then Y is not a myopically credible block of Z, which is true under either definition.

[41] To motivate weak forward induction, note that it is more difficult for agents to evaluate the credibility of proposals to change the set of contracts than the credibility of proposals that only add contracts: If a proposal changes the set of contracts, evaluating myopic credibility requires determining not just whether agents in the blocking coalition can benefit from rejecting newly proposed contracts, but also whether they might benefit from keeping their *existing* contracts. Moreover, unlike with proposals to add contracts, the credibility of a proposal to change the set of contracts does not follow from its nonstrategic individual rationality (Lemma 10).

Lemma 4 (Pareto Optimality and Forward Induction). *If there are no externalities, and $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent and satisfies Pareto optimality, it also satisfies forward induction.*

More generally, Theorem 5 shows that weak forward induction significantly refines the set of strategically consistent profiles: we can rule out all profiles from Proposition 1 except those whose stable outcomes are *maximal*. Moreover, strategically consistent profiles that satisfy weak forward induction always exist.

Theorem 5 (Weak Forward Induction: Existence and Characterization).

i. Strategically consistent profiles that satisfy weak forward induction exist.

ii. There is a strategically consistent profile satisfying weak forward induction for which $S \subseteq X$ is stable if and only if S is nonstrategically individually rational and there is no $Z \supset S$ that is nonstrategically individually rational.

Like in Theorems 2-4, the nonstrategically individually rational outcomes play an important role in Theorem 5. Because of weak forward induction, however, agents must not only believe that others will choose a nonstrategically individually rational subset from any set of available contracts Z, but a *maximal* one: That is, there can be no nonstrategically individually rational Y such that $Z \supseteq Y \supset \mu(Z)$. This is because, for blocking proposals that *add* contracts, credibility follows from nonstrategic individual rationality (Lemma 10 in the appendix).[42]

Our algorithm (3) generates beliefs with this maximality property precisely when the order used to initialize it is structured so that it ranks larger outcomes higher than smaller ones.[43] Since the resulting profile's unique stable outcome is the one that the order ranks highest, it must be maximal among all nonstrategically individually rational outcomes.

Identifying profiles that satisfy the full forward induction criterion is more challenging, precisely because it is more difficult to evaluate the credibility of blocks when they not only add new contracts, but delete existing ones. One condition that allows us to do so ensures that any beliefs that are part of a strategically consistent profile can make pairwise comparisons between nonstrategically individually rational outcomes Y, Z: i.e., either $\mu_i(Y \cup$

[42] If $Z \supset Y$ is nonstrategically individually rational, it is by definition a *myopically* credible deviation from Y. Moreover, cross-set consistency ensures that $\mu_i(Z \cup S) = S \Rightarrow \mu_i(Y \cup S) = S$, and correctness and optimality ensure that choices match beliefs (Lemma 1), so any farsighted chain starting at $Z^0 = Z$ can be changed into a farsighted chain starting at Y merely by letting $Z^0 = Y$ (farsighted credibility).

[43] If it is initialized with an order wtihout this monotonicity property, some nonstrategically individually rational outcome must be ranked below one of its subsets, and so when it is available, agents will choose — and correctly believe others will choose — only the contracts from the higher ranked subset.

$Z) = Y$ or $\mu_i(Y \cup Z) = Z$. Formally, we say that the set of outcomes $\mathcal{M} \subseteq 2^X$ is *pairwise comparable* if, for any $S, Y, Z \in \mathcal{M}$ such that $S \subseteq Y \cup Z$, either $S \subseteq Y$ or $S \subseteq Z$.

Theorem 6 (Forward Induction: Existence). *If the set of nonstrategically individually rational outcomes is pairwise comparable, a strategically consistent profile satisfying forward induction exists.*

The profile in Theorem 6 is once again constructed using our algorithm (3). To ensure that it satisfies forward induction, we initialize our algorithm with an order that not only ranks larger outcomes higher, but ensures that whenever an outcome Y is ranked above a myopically credible blocking proposal Z for Y, there is some other outcome ranked in between.[44]

When the order $\succ$ has this structure, our algorithm always generates a profile in which each agent believes that the others will go along with *all* credible blocking proposals, even those that remove existing contracts, because those proposals are always higher-ranked than the outcomes they block. In particular, given any outcome Y, any lower-ranked outcome Z is either not a farsightedly credible blocking proposal for Y (if there are outcomes ranked in between) or not a myopically credible blocking proposal for Y (otherwise).[45]

5 Discussion

Existence

Our existence results allow the application of matching-theoretic tools in new environments where agents' preferences may feature both substitutability and complementarity, agreements may have externalities, and/or more than two agents may be involved in the same contract. In particular, they employ the usual matching-theoretic approach, using a standard cooperative solution concept (stability) to pin down *outcomes* given *choice functions*. The key innovation that allows us to guarantee the existence of stable outcomes is to use *noncooperative* reasoning to determine these choice functions, thus ensuring that they are based on correct and consistent beliefs. This noncooperative reasoning cannot make predictions about outcomes on its own, because it only describes the way agents would choose from any set of contracts that might be under negotiation. Instead, applying cooperative reasoning (in the form of stability) to these choice functions makes it possible to predict the actual outcomes.

[44] The proof focuses on showing that such an order exists.

[45] When there are outcomes ranked between Z and Y, Z cannot be a farsightedly credible blocking proposal for Y: Since the nonstrategically individually rational outcomes are pairwise comparable, there is a farsighted path from Z, but not from Y, to any outcome ranked between them.

Profiles vs. Outcomes

The object pinned down by strategic consistency is not an *outcome*, but rather a profile of *choice functions*, each of which implies a unique stable outcome (Theorem 1). In particular, our results do not simply require the *set of contracts that an agent agrees to sign* to be a best response to those agreed to by others *in the stable outcome*. Rather, strategic consistency requires agents' *choice functions* to specify sets of contracts that are best responses to others' behavior *at each possible set of available contracts;* that is, at *every* set of available contracts (i.e., every set that might be discussed in a negotiation), the agents' choices must form a Nash equilibrium of a game where each agent chooses a set of contracts, and contracts go into effect if they are chosen by each agent they name.[46] Thus, a profile of strategically consistent choice functions can be interpreted as collections of equilibria of a contract-announcement game. (We formalize this connection in Section S.5 of the Online Appendix.) Stability then *selects* from among outcomes of this game by pinning down an outcome that is robust to *both individual and joint deviations* given those equilibrium choice functions. In other words, rather than being robust only to noncooperative deviations for each set of contracts, outcomes that are stable given strategically consistent behavior are robust to cooperative deviations across sets of contracts that agents evaluate strategically (i.e., given correct beliefs, rather than nonstrategically).

Beliefs and Perfection/Trembles

The beliefs involved in a strategically consistent profile might appear "brittle", in the sense that *at a given set of available contracts*, the equilibrium is not robust to trembles. In other words, an agent's choice from some set of available contracts may be weakly dominated in the contract-announcement game described in Section 3. The profile described in Part 2 of Example 1, for instance, has agent 2 choose $\emptyset$ instead of $\{x_{12}\}$ when $\{x_{12}, x_{31}\}$ are available, even though it would be better for him to choose $\{x_{12}\}$ if he thought there was a small probability that agent 1 would also choose it. And if agent 2 did choose $\{x_{12}\}$, it would be optimal for agent 1 to choose it as well.[47]

[46] This *contract-announcement game* generalizes the link-announcement game discussed in, e.g., Myerson (1991) and Jackson (2010). Strategically consistent profiles always exist precisely because this game always has a Nash equilibrium in pure strategies.

[47] Hence, at the set of available contracts $\{x_{12}, x_{31}\}$, there is no equilibrium of the contract-announcement game that avoids weakly dominated strategies and has the same outcome ($\{x_{31}\}$) as the equilibrium played there as part of this profile. At other sets of available contracts, there may be such an outcome-equivalent equilibrium; e.g., the full profile described in Footnote 22 has agent 1 choose $\emptyset$ instead of $\{x_{12}\}$ when $\{x_{12}, x_{23}\}$ are available, even though the latter weakly dominates the former in the corresponding contract-announcement game. But this produces the same outcome ($\{x_{23}\}$) as an equilibrium where agent 1 plays $\{x_{12}\}$ and agents 2 and 3 play $\{x_{23}\}$.

But if we then consider trembles at other sets of available contracts, we are left back where we started with nonstrategic choice and nonexistence. More generally, for every strategically consistent profile in Example 1, there is some set of available contracts where an agent's choice is "weakly dominated" in a way that makes it not robust to such trembles. Intuitively, even though there is a perfect equilibrium of the contract-announcement game at any *given* set of available contracts, the correct beliefs implied by these equilibria are not consistent with each other across *all* sets of available contracts. Thus, if one wishes to consider robustness to trembles along with strategic consistency, once should consider focusing on trembles that are cross-set consistent, rather than allowing all trembles independently at each set of available contracts.

Strategic vs. Nonstrategic Approach to Choice

Our approach is to work with the standard stability concept — just applied to choice functions that are *determined by a fixed point relationship with agents' beliefs about others' choices* rather than *pinned down as solutions to single-agent decision problems.* The key insight of this paper is that this allows one to make predictions in any environment, regardless of agents' preferences, externalities, or the structure of the market. But in settings where the standard, nonstrategic approach can predict stable outcomes, one might wonder (i) whether those outcomes are still stable for some strategically consistent profile, and (ii) whether there is anything to be gained by taking the strategically consistent approach in such settings instead. The answer to both questions is yes.

First, strategic consistency never overturns the predictions of the standard approach to choice, even when it is refined by weak forward induction.

Theorem 7 (Stable Outcomes with Nonstrategic vs Strategically Consistent Choice). *If an outcome is stable given the profile of nonstrategic choice functions $\{\hat{C}_i\}_{i\in I}$, it is the stable outcome of some strategically consistent assessment satisfying weak forward induction.*

Intuitively, if an outcome is stable given nonstrategic choice functions, it is, by definition, nonstrategically individually rational. Moreover, there cannot be a myopically credible blocking proposal to add contracts to it, as there must be if there was a *larger* nonstrategically individually rational outcome. Hence, it must be one of the outcomes characterized in Theorem 5.

Second, in environments where nonstrategic choice predicts a multiplicity of stable outcomes, strategic consistency allows us to attribute that multiplicity to a multiplicity of beliefs that agents may have about the way others will react to blocking proposals.[48] Each of these

[48]For an alternative interpretation of this multiplicity based on procedural fairness (the order of proposals

profiles of beliefs can be interpreted as a description of the way that bargaining power is distributed among the agents in equilibrium. And in light of Theorem 3 and Lemma 3, when we rule out coordination failure using Pareto optimality, we can regard Nash bargaining weights as sufficient statistics for the bargaining power described by beliefs.

Finally, even when stable outcomes do exist with the standard approach, strategic consistency can capture plausible outcomes that are ruled out by nonstrategic choice, a point that we illustrate with Example S.1 in the Online Appendix.[49]

Stability and Bargaining Theory

Theorem 3 can be thought of as a microfoundation for agents' beliefs by appealing to bargaining theory. If we parameterize agents' bargaining power, as in, e.g., Nash (1950), agents' beliefs, and thus (by Theorem 1) the stable outcome, can be *uniquely* pinned down. Moreover, we can identify that outcome without identifying the full profile of choice functions and beliefs. (See Section S.1 in the Online Appendix for an example.) Alternatively, with Theorem 3, we can think of strategic consistency as a (cooperative) microfoundation for a type of Nash bargaining solution in which outside options are endogenous (since the payoffs that an outcome must provide the agents in order to be individually rational depends on the payoffs they can get by dropping the agreements that they make).

6 Applications

Here, we highlight how we can use our results in three applications that have not been traditionally studied using matching-theoretic stability: network formation, bargaining with externalities, and legislative bargaining.

Network Formation

By modeling links between agents as bilateral contracts, our results can be applied to network formation settings where link formation has externalities on other agents, such as free trade agreement formation (e.g., Furusawa and Konishi (2007)) and joint venture

in a modified deferred acceptance algorithm), see Dworczak (2021).

[49]The possibility of compelling outcomes that are ruled out by stability has been noted before. In particular, a variation of stability, *weak setwise stability* (Klaus and Walzl (2009)) does not consider blocks where the participating agents' nonstrategic choices do not coincide (as in Example S.1). Formally, a blocking proposal Z is a weak setwise block of Y if each agent who participates in the block (nonstrategically) chooses the same set of contracts: for all $i \in N(Z \setminus Y)$, $\hat{C}_i(Z_i \cup Y_i | Z_{-i} \cup Y_{-i}) = Z_i$. (Such a proposal corresponds to a block $Z' = Z \setminus Y$.) Strategic consistency instead considers all blocks, but rules out disagreements about the outcome of a block by allowing agents' choices to be endogenously determined given beliefs that are correct.

formation among oligopolists (e.g., Goyal and Joshi (2003)). In these settings, arguably the most common solution concept used in the literature is *pairwise stability* (Jackson and Wolinsky (1996)), which selects networks that are robust to deviations by a pair of agents that add a link between them, and deviations by an individual agent that remove one of his links. This is a subset of the changes to the network considered by matching-theoretic stability, which also includes those in which agents substitute between links as well as those where agents add multiple links at the same time. However, pairwise stability is able to make predictions under much more general conditions (e.g., Calvó-Armengol and İlkılıç (2009)) than those known to ensure the existence of matching-theoretically stable outcomes.

But as our results show, we can ensure robustness of sets of agreements — such as the set of linking agreements that make up a network — to the full class of deviations considered by matching-theoretic stability, even without conditions on preferences or externalities. Instead of ruling out some of these deviations exogenously, strategic consistency endogenously determines which of them matter. As we illustrate in Example 4, this also allows us to make predictions in canonical network formation environments in which no pairwise stable outcome exists.

Example 4 (Trading on a Network (Jackson and Watts, 2002)). Consider the following environment from Jackson and Watts (2002). There are two divisible goods, x and y, and N consumers with identical symmetric Cobb-Douglas preferences over those goods. Before they learn their endowments, the consumers form links with one another; each link costs $c > 0$ for the two consumers that it connects. Once the network is formed, they each independently receive endowments $(1, 0)$ and $(0, 1)$ with equal probability, and trade them in separate competitive markets on each connected component of the network. Hence, adding a link benefits an agent by reducing the probability that he faces unfavorable terms of trade because most of the other agents on his connected component have the same endowment as he does.[50]

Jackson and Watts (2002) show that when $c = 5/96$ and $N \geq 4$, no pairwise stable outcome (and hence, when choice is nonstrategic, no stable outcome) exists.[51] If no agent can benefit by dropping a link, each connected component must be a tree: severing a loop does not change the equilibrium on the component, but saves c for the agents who sever their link. Moreover, if an agent has more than one link, each must be to agents that also have multiple links: otherwise, the benefit of keeping the link is lower than $5/96$. Thus, a

[50] It also reduces the probability that he faces favorable terms of trade because most agents on his connected component have the opposite endowment, but since preferences are convex, this has a smaller effect on his payoffs.

[51] As discussed in Section 6, pairwise stability is weaker than matching-theoretic stability with nonstrategic choice.

network does not provide agents with incentives to sever links — and hence, in the language of our paper, is nonstrategically individually rational — precisely when it is a collection of connected pairs. But no such network can be pairwise stable: each agent would gain more than 5/96 by forming a link with another connected pair.

Theorem 2 shows that strategic consistency resolves this nonexistence problem by requiring each agent's beliefs about how others would respond to linking proposals to be correct and consistent across different sets of available links. Moreover, Theorems 4 and 5 show that when we focus on strategically consistent profiles that satisfy either our Pareto optimality or weak forward induction refinements, matching-theoretic stability predicts exactly those outcomes we would intuitively expect to see in this setting: networks consisting of connected pairs, with no more than one isolated agent. ■

Bargaining with Externalities

Our approach also provides new insights in bargaining settings where agreements are bilateral but have externalities on other agents, such as vertical contracting (e.g., Ho and Lee (2017)) or tariff negotiations (e.g., Bagwell et al. (2020)). Solution concepts used in these settings often take an approach based on *Nash-in-Nash bargaining* (Horn and Wolinsky (1988)) — i.e., a Nash equilibrium in Nash bargains — or more explicitly based on a noncooperative alternating-offer bargaining game (e.g., Hart and Tirole (1990); Segal and Whinston (2003); Collard-Wexler et al. (2019)).[52] Both approaches commonly ensure existence and tractability by determining the outcomes of each bilateral negotiation separately, given the outcomes of the others: either directly (in the Nash-in-Nash case); by making negotiations between pairs of agents unobservable to other agents (or their delegates), even ex post, and endowing them with *passive beliefs* (e.g., Hart and Tirole (1990); De Fontenay and Gans (2014));[53] or by considering environments that satisfy conditions on the substitutability of agreements and the efficiency of the complete network (Collard-Wexler et al. (2019)).

Similarly to pairwise stability in the network formation context, this allows these concepts to offer predictions in settings where the standard approach to matching-theoretic stability does not, because agents do not (or do not need to) consider deviations that modify multiple agreements simultaneously.[54] But just as we pointed out above in the context of network

[52] As Collard-Wexler et al. (2019) show, the Nash-in-Nash approach can be microfounded as the outcome of an alternating-offer bargaining game with either delegated agents or restrictions on preferences.

[53] For an approach that does not require passive beliefs, see Segal and Whinston (2003).

[54] For instance, Nash-in-Nash bargaining does not consider, e.g., deviations to substitute between agreements with different other agents, add or remove multiple agreements with different agents at the same time, or remove some agreements with a counterparty while keeping others; matching-theoretic stability considers all of these.

formation, our results show that one can use matching-theoretic stability to find outcomes that are robust to all such deviations without making any assumptions on preferences.

Legislative Bargaining

A rich political economy literature considers settings where legislators bargain multilaterally over which of several policies to enact. Following Baron and Ferejohn (1989), this literature generally takes a noncooperative approach, modeling a negotiation as a dynamic game where legislators take turns proposing a division of surplus which is then subjected to a majoritarian vote. As Ali et al. (2019) show, the outcome of this multilateral bargaining protocol is sensitive to its extensive form, which can dramatically change the division of surplus predicted by the Baron and Ferejohn (1989) model.

Our results allow for predictions in environments where agents form multiple multilateral agreements (such as legislative bargaining) without relying on the specifics of the bargaining process. Specifically, suppose that we let contracts represent possible agreements to pass bills among (for concreteness) a majority of legislators. Then our characterization results allow us to use matching-theoretic stability to predict which of those agreements will form. For instance, when legislators' behavior is strategically consistent and satisfies the weak forward induction refinement that we introduce in Section 4.2, the legislature will pass a maximal set of bills that command the support of a majority, given the votes of the other legislators.

7 Conclusion

This paper takes a step towards the study of stable outcomes in applications where agents' preferences over agreements may exhibit both complementarities and substitutabilities, agreements can have externalities and be multilateral, and the market structure described by those agreements can be arbitrary. Our results suggest there might be new possibilities for the use of matching-theoretic models in applied work where the endogeneity of the observed agreements is of interest.

References

Agarwal, N., P. Somaini, et al. (2021): "Empirical models of non-transferable utility matching," *Online and Matching-Based Market Design*.

Ali, S. N., B. D. Bernheim, and X. Fan (2019): "Predictability and Power in Legislative Bargaining," *The Review of Economic Studies*, 86, 500–525.

Alva, S. (2018): "WARP and Combinatorial Choice," *Journal of Economic Theory*, 173, 320–333.

Aygün, O. and T. Sönmez (2012): "Matching With Contracts: The Critical Role of Irrelevance of Rejected Contracts," *Boston College Department of Economics, Working Paper*, 804.

Bagwell, K., R. W. Staiger, and A. Yurukoglu (2020): ""Nash-in-Nash" Tariff Bargaining," *Journal of International Economics*, 122, 103263.

Bando, K. (2012): "Many-to-One Matching Markets With Externalities Among Firms," *Journal of Mathematical Economics*, 48, 14–20.

Bando, K. and T. Hirai (2021): "Stability and Venture Structures in Multilateral Matching," *Journal of Economic Theory*, 105292.

Baron, D. P. and J. A. Ferejohn (1989): "Bargaining in Legislatures," *American Political Science Review*, 83, 1181–1206.

Calvó-Armengol, A. and R. İlkiliç (2009): "Pairwise-Stability and Nash Equilibria in Network Formation," *International Journal of Game Theory*, 1, 51–79.

Chakraborty, A., A. Citanna, and M. Ostrovsky (2010): "Two-sided Matching with Interdependent Values," *Journal of Economic Theory*, 145, 85–105.

Collard-Wexler, A., G. Gowrisankaran, and R. S. Lee (2019): ""Nash-in-Nash" Bargaining: A Microfoundation for Applied Work," *Journal of Political Economy*, 127, 163–195.

Crawford, G. S. and A. Yurukoglu (2012): "The Welfare Effects of Bundling in Multichannel Television Markets," *American Economic Review*, 102, 643–85.

d'Aspremont, C. and L. Gevers (2002): "Social welfare functionals and interpersonal comparability," *Handbook of social choice and welfare*, 1, 459–541.

De Fontenay, C. C. and J. S. Gans (2014): "Bilateral Bargaining with Externalities," *The Journal of Industrial Economics*, 62, 756–788.

Dworczak, P. (2021): "Deferred Acceptance with Compensation Chains," *Operations Research*, 69, 456–468.

Ellickson, B., B. Grodal, S. Scotchmer, and W. R. Zame (1999): "Clubs and the Market," *Econometrica*, 67, 1185–1217.

Fisher, J. C. and I. E. Hafalir (2016): "Matching with Aggregate Externalities," *Mathematical Social Sciences*, 81, 1–7.

Fleiner, T., R. Jagadeesan, Z. Jankó, and A. Teytelboym (2019): "Trading Networks with Frictions," *Econometrica*, 87, 1633–1661.

Furusawa, T. and H. Konishi (2007): "Free Trade Networks," *Journal of International Economics*, 72, 310–335.

Gale, D. and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69, 9–15.

Goyal, S. and S. Joshi (2003): "Networks of Collaboration in Oligopoly," *Games and Economic behavior*, 43, 57–85.

Grennan, M. (2013): "Price Discrimination and Bargaining: Empirical Evidence from Medical Devices," *American Economic Review*, 103, 145–77.

Hafalir, I. E. (2008): "Stability of Marriage with Externalities," *International Journal of Game Theory*, 37, 353–369.

Hart, O. and J. Tirole (1990): "Vertical Integration and Market Foreclosure." *Brookings Papers on Economic Activity*.

Hatfield, J. W. and F. Kojima (2008): "Matching With Contracts: Comment," *The American Economic Review*, 98, 1189–1194.

Hatfield, J. W. and S. D. Kominers (2012): "Matching in Networks with Bilateral Contracts," *American Economic Journal: Microeconomics*, 4, 176–208.

——— (2015): "Multilateral Matching," *Journal of Economic Theory*, 156, 175–206.

Hatfield, J. W., S. D. Kominers, A. Nichifor, M. Ostrovsky, and A. Westkamp (2013): "Stability and Competitive Equilibrium in Trading Networks," *Journal of Political Economy*, 121, 966–1005.

Hatfield, J. W. and P. R. Milgrom (2005): "Matching With Contracts," *American Economic Review*, 913–935.

Ho, K. and R. S. Lee (2017): "Insurer Competition in Health Care Markets," *Econometrica*, 85, 379–417.

Horn, H. and A. Wolinsky (1988): "Bilateral monopolies and incentives for merger," *The RAND Journal of Economics*, 408–419.

Jackson, M. O. (2010): *Social and Economic Networks*, Princeton University Press.

Jackson, M. O. and A. Watts (2002): "The Evolution of Social and Economic Networks," *Journal of Economic Theory*, 106, 265–295.

Jackson, M. O. and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *journal of economic theory*, 71, 44–74.

Jagadeesan, R. and K. Vocke (2021): "Stability in Large Markets," .

Kelso, A. S. and V. P. Crawford (1982): "Job Matching, Coalition Formation, and Gross Substitutes," *Econometrica: Journal of the Econometric Society*, 1483–1504.

Klaus, B. and M. Walzl (2009): "Stable many-to-many matchings with contracts," *Journal of Mathematical Economics*, 45, 422–434.

Kohlberg, E. and J.-F. Mertens (1986): "On the Strategic Stability of Equilibria," *Econometrica: Journal of the Econometric Society*, 1003–1037.

Kojima, F., P. A. Pathak, and A. E. Roth (2013): "Matching with Couples: Stability and Incentives in Large Markets," *The Quarterly Journal of Economics*, 128, 1585–1632.

Liu, C., Z. Wang, and H. Zhang (2023): "Self-Enforced Job Matching," *arXiv preprint arXiv:2308.13899*.

Liu, Q. (2020): "Stability and Bayesian Consistency in Two-Sided Markets," *American Economic Review*, 110, 2625–66.

——— (2022): "A Theory of Coalitional Games with Incomplete Information," Working paper.

Liu, Q., G. J. Mailath, A. Postlewaite, and L. Samuelson (2014): "Stable Matching with Incomplete Information," *Econometrica*, 82, 541–587.

Myerson, R. B. (1991): *Game Theory: Analysis of Conflict*, Harvard University Press.

Nash, J. F. (1950): "The Bargaining Problem," *Econometrica: Journal of the Econometric Society*, 155–162.

Nguyen, T. and R. Vohra (2018): "Near-feasible Stable Matchings with Couples," *American Economic Review*, 108, 3154–69.

Ostrovsky, M. (2008): "Stability in Supply Chain Networks," *American Economic Review*, 98, 897–923.

Pycia, M. (2012): "Stability and Preference Alignment in Matching and Coalition Formation," *Econometrica*, 80, 323–362.

Pycia, M. and M. B. Yenmez (2023): "Matching With Externalities," *The Review of Economic Studies*, 90, 948–974.

Rostek, M. and N. Yoder (2020): "Matching with Complementary Contracts," *Econometrica*, 88, 1793–1827.

——— (2023): "Strategic Consistency in Two-Sided Matching Markets," Working Paper.

——— (2025): "Counterfactual Analysis in Bargaining with Externalities: A Matching-Theoretic Foundation," Working paper.

Sadler, E. (2023): "A Unified Approach to Strategic Network Formation and Classical Matching Theory," *Available at SSRN*.

Sasaki, H. and M. Toda (1996): "Two-Sided Matching Problems with Externalities," *Journal of Economic Theory*, 70, 93–108.

Saulle, R., P. Salmaso, and A. Nicolò (2025): "Rationalizable Behavior in Matching with Externalities," *Available at SSRN 5874622*.

Segal, I. and M. D. Whinston (2003): "Robust Predictions for Bilateral Contracting with Externalities," *Econometrica*, 71, 757–791.

Sun, N. and Z. Yang (2006): "Equilibria and Indivisibilities: Gross Substitutes and Complements," *Econometrica*, 74, 1385–1402.

——— (2009): "A Double-Track Adjustment Process for Discrete Markets With Substitutes and Complements," *Econometrica*, 77, 933–952.

Teytelboym, A. (2014): "Gross Substitutes and Complements: A Simple Generalization," *Economics Letters*, 123, 135–138.

Appendix

Proof of Lemma 1 For each $Y \subseteq X$, let $C(Y) := \cap_{i \in I}(C_i(Y_i|Y_{-i}) \cup Y_{-i})$. For each $i \in I$ and $Y \subseteq X$, since C_i is optimal given μ_i, we have $C_i(Y_i|Y_{-i}) \subseteq \mu_i(Y)$, and since μ_i is correct

given $\{C_j\}_{j\neq i}$, we have $\mu_i(Y) = C_{-i}(Y)$. Then for each $i \in I$, we have $C_i(Y_i|Y_{-i}) \subseteq C_{-i}(Y)$, and hence

$$C_i(Y_i|Y_{-i}) = C_i(Y_i|Y_{-i}) \cap C_{-i}(Y) = ((C_i(Y_i|Y_{-i}) \cup Y_{-i}) \cap X_i) \cap C_{-i}(Y) = X_i \cap C(Y). \quad (6)$$

Then for each $j \in I$ and $Y \subseteq X$, we have $C_{-j}(Y) = C(Y)$: By definition, $C(Y) = (C_j(Y_j|Y_{-j}) \cup Y_{-j}) \cap C_{-j}(Y)$, so $C(Y) \subseteq C_{-j}(Y)$. Now suppose $x \in C_{-j}(Y)$. By assumption, $|N(x)| \geq 2$, so we must have $x \in C_i(Y_i|Y_{-i})$ for some $i \neq j$. Since $C_i(Y_i|Y_{-i}) = X_i \cap C(Y) \subseteq C(Y)$, it follows that $C_{-j}(Y) \subseteq C(Y)$.

Then since beliefs are correct, for all $i \in I$ and $Y \subseteq X$, we have $\mu_i(Y) = C(Y)$; (i) follows for $\mu(Y) = C(Y)$, and thus (ii) follows from (6). □

Lemma 5 adds to Lemma 1 by showing that given correctness and optimality, cross-set consistency is equivalent to the weak axiom (or equivalently, the irrelevance of rejected contracts condition (Alva (2018))) on the agents' common beliefs.

Lemma 5. *Suppose that the choice functions $\{C_i\}_{i\in I}$ are optimal given beliefs $\{\mu_i\}_{i\in I}$, and the beliefs $\{\mu_i\}_{i\in I}$ are correct given choice functions $\{C_i\}_{i\in I}$. Then $\{\mu_i\}_{i\in I}$ are cross-set consistent given $\{C_i\}_{i\in I}$ if and only if for each $i \in I$, $Y \supseteq Z \supseteq \mu_i(Y)$ implies $\mu_i(Y) = \mu_i(Z)$.*

Proof. By Lemma 1, for each $i, j \in I$ and $Y \subseteq X$, we have $\mu_i(Y) = \mu_j(Y) = \mu(Y)$ and $C_j(Y_j|Y_{-j}) = \mu(Y) \cap X_j$. Hence, for each $Y, Z \subseteq X$, it holds that $Y \supseteq Z \supseteq C_j(Y_j|Y_{-j})$ for each $j \in I$ if and only if $Y \supseteq Z \supseteq (\bigcup_{j\in I} \mu(Y) \cap X_j) = \mu(Y)$. Then since $\mu_i(S) = \mu_j(S) = \mu(S)$ for each $i, j \in I$ and $S \subseteq X$, it follows that $\{\mu_i\}_{i\in I}$ are cross-set consistent given $\{C_i\}_{i\in I}$ if and only if for each $Y, Z \subseteq X$ such that $Y \supseteq Z \supseteq \mu(Y)$, we have $\mu(Y) = \mu(Z)$. □

Proof of Theorem 1 (Strategic Consistency and Stability) Let $\{C_i, \mu_i\}_{i\in I}$ be a strategically consistent profile. By Lemma 1, for each $i, j \in I$ and $S \subseteq X$, we have $\mu_i(S) = \mu_j(S) = \mu(S)$ and $C_j(S_j|S_{-j}) = \mu(S) \cap X_j$.

Now for any $Z \subseteq X$, we have $X \supseteq \mu(X) \cup Z \supseteq \mu(X)$. Then by Lemma 5, $\mu(\mu(X) \cup Z) = \mu(X)$. Then by Lemma 1 (ii), for each $Z \subseteq X \setminus \mu(X)$ and each $i \in I$, we have $C_i((\mu(X) \cup Z)_i|(\mu(X) \cup Z)_{-i}) = \mu(\mu(X) \cup Z) \cap X_i = \mu(X) \cap X_i$. It follows that $\mu(X)$ is unblocked and (by setting $Z = \emptyset$) individually rational. □

Proof of Corollary 1 (Stability and Beliefs) Follows immediately from the proof of Theorem 1. □

Lemma 6 shows that optimal choices from Y given beliefs μ_i (as in a strategically consistent profile) are the same as nonstrategic choices from the set of contracts $\mu_i(Y)$ that an agent believes the other agents will choose from Y.

Lemma 6. *C_i is optimal given μ_i if and only if $C_i(Y_i|Y_{-i}) = \hat{C}_i(\mu_i(Y) \cap X_i|\mu_i(Y) \cap X_{-i})$ for all $Y \subseteq X$.*

Proof. From (1), we have $\hat{C}_i(\mu_i(Y)_i|\mu_i(Y)_{-i}) = \arg\max_{S \subseteq \mu_i(Y)_i} u_i(S \cup \mu_i(Y)_{-i})$; the statement follows immediately from the definition of optimality of C_i given μ_i. □

Proof of Lemma 2 (Converse of Lemma 1) (a): From condition (ii), for each $i \in I$ and $Y \subseteq X$,

$$C_{-i}(Y) := \bigcap_{j \neq i} (C_j(Y_j|Y_{-j}) \cup Y_{-j}) = \bigcap_{j \neq i} ((\mu_i(Y) \cap X_j) \cup Y_{-j}). \tag{7}$$

Then $\mu_i(Y) = C_{-i}(Y)$: If $x \in C_{-i}(Y)$, then since $|N(x)| \geq 2$, we must have $x \in X_j$ for some $j \neq i$, and hence, by (7), $x \in \mu_i(Y) \cap X_j \subseteq \mu_i(Y)$; it follows that $\mu_i(Y) \supseteq C_{-i}(Y)$. Moreover, since $\mu_i(Y) \subseteq Y$, we have $\mu_i(Y) \subseteq ((\mu_i(Y) \cap X_j) \cup Y_{-j})$ for each j, and so by (7), $\mu_i(Y) \subseteq C_{-i}(Y)$. Hence, $\{\mu_i\}_{i \in I}$ are correct given $\{C_i\}_{i \in I}$.

(b): (Beliefs are nonstrategically IR ⇒ Choices are optimal) Suppose that $\mu(Y)$ is nonstrategically individually rational for each $Y \subseteq X$. Then by definition, for each $i \in I$ and $Y \subseteq X$, we have $\hat{C}_i(\mu(Y) \cap X_i|\mu(Y) \cap X_{-i}) = \mu(Y) \cap X_i = C_i(Y_i|Y_{-i})$. It follows from Lemma 6 that $\{C_i\}_{i \in I}$ are optimal given $\{\mu_i\}_{i \in I}$.

(Choices are optimal ⇒ Beliefs are nonstrategically IR) Suppose that $\{C_i\}_{i \in I}$ are optimal given $\{\mu_i\}_{i \in I}$. For each $Y \subseteq X$ and $i \in I$, we have $\mu_i(Y) \cap X_i = C_i(Y_i|Y_{-i})$ (by condition (ii)) and $C_i(Y_i|Y_{-i}) = \hat{C}_i(\mu(Y) \cap X_i|\mu(Y) \cap X_{-i})$ (by Lemma 6). Then for each $Y \subseteq X$ and $i \in I$, we have $\mu(Y) \cap X_i = \hat{C}_i(\mu(Y) \cap X_i|\mu(Y) \cap X_{-i})$, and so $\mu(Y)$ is nonstrategically individually rational. □

Lemma 7 shows that our algorithm (3) always constructs a strategically consistent profile.

Lemma 7 (Construction Algorithm). *For any strict total order $\succ$ on the set $\mathcal{M} = \{Y \subseteq X|\hat{C}_i(Y_i|Y_{-i}) = Y_i \text{ for each } i \in I\}$ of nonstrategically individually rational outcomes, the profile of choice functions and beliefs $\{C_i, \mu_i\}_{i \in I}$ defined in (3) is strategically consistent.*

Proof. First note that $\{C_i, \mu_i\}_{i \in I}$ is well-defined: Since $\{\hat{C}_i\}_{i \in I}$ are nonstrategic, we have $\hat{C}_i(\emptyset|\emptyset) = \emptyset$ for each $i \in I$, and so $\emptyset \in \mathcal{M}$. Then for each $Y \subseteq X$, the collection $\{Y'|Y' \subseteq Y\}$ contains at least one element of $\mathcal{M}$, and so $\mu(Y) = \max_{\succ}\{Y'|Y' \subseteq Y\}$ is well-defined for each Y.

By construction, for each $i, j \in I$ and $Y \subseteq X$, (i) $\mu_i(Y) = \mu_j(Y) := \mu(Y)$, and (ii) $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$. Then by Lemma 2 (a), $\{\mu_i\}_{i \in I}$ are correct given $\{C_i\}_{i \in I}$. And since $\mu_i(Y) \in \mathcal{M}$ for each $Y \subseteq X$, by Lemma 2 (b), $\{C_i\}_{i \in I}$ are optimal given $\{\mu_i\}_{i \in I}$.

To show cross-set consistency, suppose $\mu(Y) \subseteq Z \subseteq Y$ for some $i \in I$ and $Y, Z \subseteq X$. Then by construction, $\mu(Y) \succeq Y'$ for each $Y' \subseteq Y$, and hence for each $Y' \subseteq Z$. Then by construction, $\mu(Y) = \mu(Z)$. It follows from Lemma 5 that $\{\mu_i\}_{i \in I}$ are cross-set consistent given $\{C_i\}_{i \in I}$. □

Lemma 8 (Strategically Consistent Profiles: Characterization). *There is a strategically consistent profile for which the outcome $Y \subseteq X$ is stable if and only if Y is nonstrategically individually rational.*

Proof. (Only if) Suppose S is stable for the strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$. By Corollary 1, $S = \mu_i(X)$ for each $i \in I$. Since $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent, by Lemma 1, for all $i, \in I$ and $Y \subseteq X$, we have $\mu_i(Y) = \mu_j(Y) = \mu(Y)$ and $C_i(Y_i|Y_{-i}) = \mu(Y) \cap X_i$. Then by Lemma 2 (b), S is nonstrategically individually rational.

(If) Suppose S is nonstrategically individually rational. Choose any strict total order $\succ$ on the set $\mathcal{M} = \{Y \subseteq X | \hat{C}_i(Y_i|Y_{-i}) = Y_i \text{ for each } i \in I\}$ of nonstrategically individually rational outcomes which ranks S highest. Let $\{C_i, \mu_i\}_{i \in I}$ be the profile of choice functions and beliefs constructed according to (3). By Lemma 7, $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent. Since $S \succeq Y$ for all $Y \in \mathcal{M}$, it follows that $S = \max_{\succ}\{Y' | Y' \subseteq X\} = \mu_i(X)$ for each $i \in I$. Then by Corollary 1, $\mu(X) = S$ is uniquely stable for $\{C_i, \mu_i\}_{i \in I}$. □

Proof of Theorem 2 (Existence of Strategically Consistent Profiles) By definition, $\hat{C}_i(\emptyset|\emptyset) = \emptyset$ for each $i \in I$. Hence, $\emptyset$ is nonstrategically individually rational; existence follows from the "if" part of Lemma 8. □

Lemma 9. *If $\phi : \mathbb{R}^I_+ \to \mathbb{R}$ is a strictly increasing function, then for any $Y \in \arg\max_{S \in \mathcal{M}} \phi((u_i(S))_{i \in I})$, there is a strict total order $\succ^\phi$ that is induced by ϕ which ranks Y highest.*

Proof. Let $\mathcal{M} = \{Z \subseteq X | \hat{C}_i(Z_i|Z_{-i}) = Z_i \text{ for each } i \in I\}$ denote the set of nonstrategically individually rational outcomes, and label its elements according to the sequence $\{Y^n\}_{n=1}^{|\mathcal{M}|}$, constructed recursively as follows: To begin, let $Y^1 = Y \in \arg\max_{S \in \mathcal{M}} \phi((u_i(S))_{i \in I})$. Then, given elements $\{Y^n\}_{n=1}^m$, choose $Y^{m+1} \in \arg\max_{S \in \mathcal{M} \setminus \{Y^n\}_{n=1}^m} \phi((u_i(S))_{i \in I})$. (The set of maximizers is nonempty since X (and hence $\mathcal{M} \subseteq 2^X$) is finite.) This construction implies that whenever $\phi((u_i(Y^n))_{i \in I}) > \phi((u_i(Y^m))_{i \in I})$, we must have $n < m$: If $n > m$, then $Y^n \in \mathcal{M} \setminus \{Y^k\}_{k=1}^{m-1}$, and so Y^m could not have been chosen as the mth element of the sequence.

Now define the order $\succ^\phi$ on $\mathcal{M}$ as follows: $Y^n \succ^\phi Y^m \Leftrightarrow n < m$. Since $\{Y^n\}_{n=1}^m = \mathcal{M}$, we can label any two elements of $\mathcal{M}$ as Y^n and Y^m for some n, m. If $\phi((u_i(Y^n))_{i \in I}) > \phi((u_i(Y^m))_{i \in I})$, we must have $n < m$, and hence $Y^n \succ^\phi Y^m$. So $\succ^\phi$ is induced by ϕ, as desired. □

Proof of Theorem 3 (Pareto-Optimal Profiles) (i): Strategic consistency follows from Lemma 7. For Pareto optimality, suppose that $Y, Z \subseteq X$ are such that $u_i(Y) \geq u_i(Z)$ for all $i \in I$ and $u_i(Y) > u_i(Z)$ for some $i \in I$. Then since ϕ is strictly increasing, we must have $\phi((u_i(Y))_{i\in I}) > \phi((u_i(Z))_{i\in I})$. Thus, since $\succ^\phi$ is induced by ϕ, we have $Y \succ^\phi Z$. Then the algorithm (3) yields $\mu_i^\phi(Y \cup Z) = \mu^\phi(Y \cup Z) = \max_{\succ^\phi}\{Y' | Y' \subseteq Y \cup Z\} \neq Z$ for each $i \in I$. It follows that $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$ satisfies Pareto optimality.

(ii): For each $i \in I$ and $Z \subseteq X$, we have $\mu_i^\phi(Z) = \mu^\phi(Z) = \max_{\succ^\phi}\{S | S \subseteq Z\}$ from (3). Then $\mu_i^\phi(Z) \succ^\phi S$ for all $S \subseteq Z$. Since $\succ^\phi$ is induced by ϕ, it follows that for all $S \subseteq Z$, we have $\phi((u_j(\mu_i^\phi(Z)))_{j\in I}) \geq \phi((u_j(S))_{j\in I})$: If not, then $\phi((u_j(\mu_i^\phi(Z)))_{j\in I}) < \phi((u_j(S))_{j\in I})$, and hence $\mu_i^\phi(Z) \prec^\phi S$. It follows that $\mu_i^\phi(Z)$ solves (4). □

Proof of Theorem 4 (Welfare Theorem for Strategic Consistency) (Only if) Suppose that Y is stable for a strategically consistent profile $\{C_i, \mu_i\}_{i\in I}$ that satisfies Pareto optimality. By Lemma 8, Y is nonstrategically individually rational. Suppose there is another nonstrategically individually rational outcome Z that Pareto-dominates Y: $u_i(Z) \geq u_i(Y)$ for all $i \in I$ and $u_i(Z) > u_i(Y)$ for some $i \in I$. Then since $\{C_i, \mu_i\}_{i\in I}$ satisfies Pareto optimality, $\mu_i(Z \cup Y) = Z$ for each $i \in I$. Then we must have $\mu_i(X) \neq Y$ for each $i \in I$; otherwise, cross-set consistency would imply $\mu_i(Z \cup Y) = Y \neq Z$. Then by Corollary 1, Y is not stable for $\{C_i, \mu_i\}_{i\in I}$, a contradiction.

(If) Denote by $\mathcal{M}$ the nonstrategically individually rational outcomes, and suppose that Y is Pareto efficient among these outcomes: there exists no $S \in \mathcal{M}$ such that $u_i(S) \geq u_i(Y)$ for all $i \in I$ and $u_i(S) > u_i(Y)$ for some $i \in I$. Then for every $S \in \mathcal{M}$, either $u_i(S) = u_i(Y)$ for all $i \in I$ or $u_i(S) < u_i(Y)$ for some $i \in I$. For each $\rho < 0$, define

$$\begin{aligned} \phi_\rho : \mathbb{R}_+^I &\to \mathbb{R} \\ x &\mapsto \left(\sum_{i\in I}\left(\frac{x_i}{u_i(Y)}\right)^\rho\right)^{1/\rho}. \end{aligned}$$

Each ϕ_ρ is strictly increasing, since

$$\frac{\partial \phi_\rho}{\partial x_i}(x) = \frac{x_i^{\rho-1}}{u_i(Y)^\rho}\left(\sum_{i\in I}\left(\frac{x_i}{u_i(Y)}\right)^\rho\right)^{1/\rho-1} > 0 \text{ for each } i \in I.$$

Now for any $Z \in \mathcal{M}$, we have

$$\begin{aligned}\phi_\rho((u_i(Y))_{i\in I}) - \phi_\rho((u_i(S))_{i\in I}) &= 1 - \left(\sum_{i\in I}\left(\frac{u_i(S)}{u_i(Y)}\right)^\rho\right)^{1/\rho};\\ \lim_{\rho\to-\infty}\left(\phi_\rho((u_i(Y))_{i\in I}) - \phi_\rho((u_i(S))_{i\in I})\right) &= 1 - \min\left\{u_i(S)/u_i(Y)\right\}_{i\in I}\\ &< 0, \text{ if } (u_i(S))_{i\in I} \neq (u_i(Y))_{i\in I}.\end{aligned}$$

Then for every $S \in \mathcal{M}$ with $(u_i(Y))_{i\in I} \neq (u_i(S))_{i\in I}$, there exists r_S such that for all $\rho < r_S$, $\phi_\rho((u_i(Y))_{i\in I}) > \phi_\rho((u_i(S))_{i\in I})$. Choose $\rho^* = \min_{S\in\mathcal{M}} r_S$ and let $\phi = \phi_{\rho^*}$; it follows that $Y \in \arg\max_{S\in\mathcal{M}} \phi((u_i(S))_{i\in I})$. Then by Lemma 9, there is a strict total order $\succ^\phi$ that is induced by ϕ and ranks Y highest, and by Theorem 3(i), the profile $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$ constructed from $\succ^\phi$ using the algorithm (3) is strategically consistent and satisfies Pareto optimality. By (3), since $\succ^\phi$ ranks Y highest, $\mu_i(X) = Y$ for each $i \in I$; it follows from Corollary 1 that Y is stable for $\{C_i^\phi, \mu_i^\phi\}_{i\in I}$, as desired. □

Proof of Lemma 3 Follows from d'Aspremont and Gevers (2002) Theorem 4.17. □

Proof of Lemma 4 (Pareto Optimality and Forward Induction) Suppose that given $\{C_i, \mu_i\}_{i\in I}$, Z is a credible blocking proposal for Y. Fix a nonstrategically individually rational $S \subseteq Y \cup Z$ with $S \neq Z$. Since there are no externalities, myopic credibility implies $u_i(Z_i) \geq u_i(S_i)$ for each $i \in I$. By assumption, $u_i(Z_i) \neq u_i(S_i)$, and hence $u_i(Z_i) > u_i(S_i)$, for each $i \in I$ such that $S_i \neq Z_i$. Then since $\{C_i, \mu_i\}_{i\in I}$ satisfies Pareto optimality, and Z is nonstrategically individually rational, we have $\mu_i(Z \cup S) \neq S$ for each $i \in I$. Then by cross-set consistency, we have $\mu_i(Z \cup Y) \neq S$ for each $i \in I$.

By Lemmas 1 and 2 (b), for each $i \in I$, we have that $\mu_i(Z \cup Y)$ is nonstrategically individually rational. Then by elimination, we must have $\mu_i(Z \cup Y) = Z$ for each $i \in I$. Thus, $\{C_i, \mu_i\}_{i\in I}$ satisfies forward induction. □

Lemma 10. *Given a strategically consistent profile $\{C_i, \mu_i\}_{i\in I}$, $Z \supset Y$ is a credible blocking proposal for Y if and only if it is nonstrategically individually rational.*

Proof. (If) Suppose $Z \supset Y$ is nonstrategically individually rational.

Z is a myopically credible blocking proposal for Y: Since $Z \supset Y$ and Z is nonstrategically individually rational, $Z_i = \hat{C}_i(Z_i|Z_{-i}) = \hat{C}_i((Y \cup Z)_i|(Y \cup Z)_{-i})$ for each $i \in I$.

Z is a farsightedly credible blocking proposal for Y: Suppose there is a farsighted chain $\{Z^n\}_{n=0}^N$ from Z to Z'. Then for each $i \in I$, we have $\mu_i(Z \cup Z^1) = Z^1$. Then since $\{C_i, \mu_i\}_{i\in I}$ is strategically consistent and $Z \supset Y$, by Lemma 5, $\mu_i(Y \cup Z^1) = Z^1$. Then $\{Y, \{Z^n\}_{n=1}^N\}$ is a farsighted chain from Y to Z'.

Then since Z is nonstrategically individually rational, it is a credible blocking proposal for Y.

(Only if) Suppose $Z \supset Y$ is a credible blocking proposal for Y. Then by definition, it is nonstrategically individually rational. □

Corollary 2. *A strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$ satisfies weak forward induction if and only if $\mu_i(Y) = \mu(Y) = Y$ for each $i \in I$ whenever Y is nonstrategically individually rational.*

Proof. (Only if) Suppose $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent and satisfies weak forward induction, and that Y is nonstrategically individually rational. If $Y = \emptyset$, then $C_i(Y_i|Y_{-i}) = \emptyset$ for each $i \in I$; then since beliefs are correct, $\mu_i(Y) = \emptyset = Y$ for each $i \in I$. Alternatively, if $Y \neq \emptyset$, then by Lemma 10, Y is a credible blocking proposal for $\emptyset$. Moreover, $\emptyset$ is nonstrategically individually rational, since by definition, $\hat{C}_i(\emptyset|\emptyset) = \emptyset$ for each $i \in I$. Then since $\{C_i, \mu_i\}_{i \in I}$ satisfies weak forward induction, we must have $\mu_i(Y) = \mu_i(Y \cup \emptyset) = Y$ for each $i \in I$.

(If) Suppose $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent and that $\mu_i(Y) = Y$ for each nonstrategically individually rational $Y \subseteq X$ and each $i \in I$. Consider $Y, Z \subseteq X$ such that Z is a credible blocking proposal for Y, and $Z \supset Y$. By definition, Z is nonstrategically individually rational; then $\mu_i(Y \cup Z) = \mu_i(Z) = Z$. Then $\{C_i, \mu_i\}_{i \in I}$ satisfies weak forward induction. □

Proof of Theorem 5 (Weak Forward Induction: Existence and Characterization) (ii): (Only if) Suppose that S is stable for the strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$, and that $\{C_i, \mu_i\}_{i \in I}$ satisfies weak forward induction. Then by Lemma 8, S is nonstrategically individually rational. Now suppose toward a contradiction that there exists $Z \supset S$ that is also nonstrategically individually rational. Then by Corollary 2, $\mu_i(Z) = Z$ for each $i \in I$, since $\{C_i, \mu_i\}_{i \in I}$ satisfies weak forward induction. Since S is stable for $\{C_i, \mu_i\}_{i \in I}$, by Corollary 1, $\mu_i(X) = S$ for each $i \in I$. Then since $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent, by Lemma 5, $\mu_i(Z) = S$ for each $i \in I$, a contradiction.

(If) Suppose that S is nonstrategically individually rational and there is no $Z \supset S$ that is nonstrategically individually rational. To construct a strategically consistent profile $\{C_i, \mu_i\}_{i \in I}$ satisfying WFI for which S is stable, we first define an order $\succ$ that refines the subset order $\supset$ and ranks S highest, and then use it in our algorithm (3) and show that the resulting profile satisfies weak forward induction.

Let $\mathcal{M} = \{Z \subseteq X | \hat{C}_i(Z_i|Z_{-i}) = Z_i \text{ for each } i \in I\}$ denote the set of nonstrategically individually rational outcomes, and label its elements according to the sequence $\{Y^n\}_{n=1}^{|\mathcal{M}|}$, constructed inductively as follows: For the initial element, choose $Y^1 = S$. Then, given

elements $\{Y^n\}_{n=1}^m$, choose $Y^{m+1} \in \mathcal{M} \setminus \{Y^n\}_{n=1}^m$ such that there is no $Y' \in \mathcal{M} \setminus \{Y^n\}_{n=1}^m$ with $Y' \supset Y^{m+1}$. This construction implies that whenever $Y^n \supset Y^m$, we must have $n < m$: If $n > m$, then $Y^n \in \mathcal{M} \setminus \{Y^k\}_{k=1}^{m-1}$, and so Y^m could not have been chosen as the mth element of the sequence.

Now define the order $\succ$ on $\mathcal{M}$ as follows: $Y^n \succ Y^m \Leftrightarrow n < m$. This order refines $\supset$ on $\mathcal{M}$: If $Y^n \supset Y^m$, then $n < m$, and hence $Y^n \succ Y^m$.

Let $\{C_i, \mu_i\}_{i \in I}$ be the profile of choice functions and beliefs constructed according to (3). By Lemma 7, $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent. Since $S = Y^1$, we have $S \succeq Y$ for all $Y \in \mathcal{M}$, and hence $S = \max_{\succ}\{Y' | Y' \subseteq X\} = \mu_i(X)$ for each $i \in I$. Then by Corollary 1, $\mu(X) = S$ is uniquely stable for $\{C_i, \mu_i\}_{i \in I}$.

Now we show that $\{C_i, \mu_i\}_{i \in I}$ satisfies weak forward induction. Since $\succ$ refines $\supset$ on $\mathcal{M}$, for any $Y, Y' \in \mathcal{M}$ with $Y \supset Y'$, we have $Y \succ Y'$. Then by (3), if $Y \in \mathcal{M}$, $\mu_i(Y) = \max_{\succ}\{Y' | Y' \subseteq Y\} = Y$ for each $i \in I$. It follows from Corollary 2 that $\{C_i, \mu_i\}_{i \in I}$ satisfies weak forward induction.

(i): The set $\mathcal{M}$ of nonstrategically individually rational outcomes is nonempty, since it always contains the autarky outcome $\emptyset$. (By definition, we must have $\hat{C}_i(\emptyset|\emptyset) = \emptyset$ for all $i \in I$.) Moreover, since X is finite, so is 2^X, and so $\mathcal{M} \subseteq 2^X$ is finite as well. It follows straightforwardly that $\mathcal{M}$ must contain at least one element S such that there is no $Z \in \mathcal{M}$ with $Z \supset S$. □

The proof of Theorem 6 relies on techniques from graph theory.[55] Consider a directed graph which has an edge from Z to Y whenever Z is *not* a myopically credible blocking proposal for Y. Then, consider the set of paths on this graph that (a) are $\subseteq$-nondecreasing, i.e., never pass through Z after $Y \supset Z$, and (b) do not pass through any nodes more than once. If we choose one of the longest of these paths $\{Y^n\}_{n=1}^N$, it must pass through each node: Any node Y the path does not pass through can be inserted somewhere along the path that is after each of its subsets $Z \subset Y$ and before each of its supersets $S \supset Y$.

Then we can construct the profile in Theorem 6 the same way (3) as in Theorem 2, but this time ordering the nonstrategically individually rational outcomes by their position along the path: $Y \succ Z \Leftrightarrow Y = Y^n$ and $Z = Y^m$ for $n > m$. Because of pairwise comparability and the $\subseteq$-nondecreasing nature of the path, agents always believe that a higher-ranked outcome will result from a block of a lower-ranked outcome: $Y \succ Z \Rightarrow \mu(Y \cup Z) = Y$.

Consequently, forward induction only requires that lower-ranked outcomes Y^m are not credible blocking proposals for higher-ranked ones Y^n. If the higher-ranked one is the direct successor of the lower-ranked one — i.e., if $n = m+1$ — then by the path's construction, it is not a myopically credible blocking proposal. Otherwise, if $n > m+1$, there is some outcome

[55] We thank Akhil Vohra for the idea for this construction.

Y^k ranked between the other two — i.e., with $n > k > m$ — which is (by construction) at the end of a farsighted chain from the lower-ranked outcome, but not the higher-ranked one.

Proof of Theorem 6 (Forward Induction: Existence)

Step 0: The myopically non-credible order $\trianglerighteq$. Let $\mathcal{M} = \{Z \subseteq X | \hat{C}_i(Z_i|Z_{-i}) = Z_i \text{ for each } i \in I\}$ denote the set of nonstrategically individually rational outcomes, and define an order $\trianglerighteq$ on $\mathcal{M}$ as follows: $Y \trianglerighteq Z \Leftrightarrow$ either Z is *not* a myopically credible blocking proposal for Y (i.e., $Z_i \neq \hat{C}_i((Y \cup Z)_i|(Y \cup Z)_{-i})$ for some $i \in I$) or $Z = Y$.

$\trianglerighteq$ is complete: Suppose toward a contradiction that $Y \not\trianglerighteq Z$ and $Z \not\trianglerighteq Y$. Then $Z_i = \hat{C}_i((Y \cup Z)_i|(Y \cup Z)_{-i}) = Y_i$ for all $i \in I$. Then since $N(x) \neq \emptyset$ for each x, we have $Z = \bigcup_{i \in I} Z_i = \bigcup_{i \in I} Y_i = Y$, a contradiction.

Moreover, by Lemma 10, $\triangleright$ refines $\supset$: if $Z \supset Y$, then $Z \triangleright Y$. Hence, by contrapositive, $Z \trianglelefteq Y \Rightarrow Z \not\supset Y$.

Step 1: Choose a longest $\supset$-nondecreasing, nonrepeating, $\trianglerighteq$-successor path $\{Z^n\}_{n=1}^M$. Let $\mathcal{Y}$ be the set of sequences $\{Y^n\}_{n=1}^N \subseteq \mathcal{M}$ such that (a) $n < m$ whenever $Y^n \supset Y^m$, (b) $Y^n \neq Y^m$ whenever $n \neq m$, and (c) for each n, $Y^n \trianglerighteq Y^{n+1}$. Since X is finite, so is 2^X, and hence $\mathcal{M}$. From property (b), any sequence in $\mathcal{Y}$ can have at most $|\mathcal{M}|$ elements from the finite set $\mathcal{M}$, so $\mathcal{Y}$ is finite as well. Then it must have a longest element $\{Z^n\}_{n=1}^M$ such that for any $\{Y^n\}_{n=1}^N \in \mathcal{Y}$, we have $N \leq M$.

Step 2: The path terminates at $Z^M = \emptyset$. Observe that $\emptyset \in \mathcal{M}$, since by definition, $\hat{C}_i(\emptyset|\emptyset) = \emptyset$ for each $i \in I$. $\emptyset \subseteq Z^n$ for all n, so by condition (a), either $\emptyset = Z^M$ or $\emptyset \notin \{Z^n\}_{n=1}^M$. Suppose toward a contradiction that the latter is true. Then we can append $\emptyset$ to $\{Z^n\}_{n=1}^M$ to create a longer sequence $\{\{Z^n\}_{n=1}^M, \emptyset\}$ that is still part of $\mathcal{Y}$: (a) holds since it holds for $\{Z^n\}_{n=1}^M$, (b) holds since it holds for $\{Z^n\}_{n=1}^M$ and $\emptyset \notin \{Z^n\}_{n=1}^M$, and (c) holds since it holds for $\{Z^n\}_{n=1}^M$ and (since $\triangleright$ refines $\supset$ and $Z^M \supset \emptyset$) we have $Z^M \triangleright \emptyset$. But $\{Z^n\}_{n=1}^M$ is the longest sequence in $\mathcal{Y}$, a contradiction.

Step 3: The path covers all of $\mathcal{M}$: For each $Y \in \mathcal{M}$, $Y \in \{Z^n\}_{n=1}^M$. Suppose toward a contradiction that $Y \notin \{Z^n\}_{n=1}^M$, and let $K = \min\{n | Z^n \subset Y\}$. ($\{n|Z^n \subset Y\}$ is nonempty, since $Z^M = \emptyset \subset Y$ by Step 2.) Then $Z^K \subset Y$, and since $\triangleright$ refines $\supset$, $Y \triangleright Z^K$.

By definition, $Z^n \subset Y \Rightarrow n \geq K$. Moreover, $Z^n \supset Y \Rightarrow n < K$: if $Z^n \supset Y$, then $Z^n \supset Z^K$, and so by (a) $n < K$.

We use induction to show that $Y \triangleright Z^n$ (and hence, since $\triangleright$ refines $\supset$, we have $Y \not\subset Z^n$) for all $n < K$:

- **Initial step: $Y \triangleright Z^{K-1}$.** Suppose toward a contradiction that $Y \not\triangleright Z^{K-1}$. Since $\trianglerighteq$ is complete, $Y \trianglelefteq Z^{K-1}$. Then $\{\{Z^n\}_{n=1}^{K-1}, Y, \{Z^n\}_{n=K}^M\} \in \mathcal{Y}$: (c) is satisfied since it holds for $\{Z^n\}_{n=1}^M$ and $Z^{K-1} \trianglerighteq Y \triangleright Z^K$. (b) holds since it holds for $\{Z^n\}_{n=1}^M$ and $Y \notin \{Z^n\}_{n=1}^M$.

Finally, (a) holds since it holds for $\{Z^n\}_{n=1}^M$, and we know that $Z^n \subset Y \Rightarrow n \geq K$, and $Z^n \supset Y \Rightarrow n < K$. But $\{Z^n\}_{n=1}^M$ is the longest sequence in $\mathcal{Y}$, a contradiction.

- **Induction step: for any $t \leq K-1$, if $Y \triangleright Z^n$ for all $n \in [t, K-1]$, then $Y \triangleright Z^n$ for all $n \in [t-1, K-1]$.** Suppose that $Y \triangleright Z^n$ for all $n \in [t, K-1]$. Since $\triangleright$ refines $\supset$, it follows that $Z^n \not\supset Y$ for all $n \in [t, K-1]$. Since we have already shown that $Z^n \supset Y \Rightarrow n < K$, it follows that $Z^n \supset Y \Rightarrow n < t$.

  Now suppose toward a contradiction that $Y \not\trianglerighteq Z^{t-1}$. Since $\trianglerighteq$ is complete, $Y \trianglelefteq Z^{t-1}$. Then $\{\{Z^n\}_{n=1}^{t-1}, Y, \{Z^n\}_{n=t}^M\} \in \mathcal{Y}$: (c) is satisfied since it holds for $\{Z^n\}_{n=1}^M$ and $Z^{t-1} \trianglerighteq Y \triangleright Z^t$. (b) holds since it holds for $\{Z^n\}_{n=1}^M$ and $Y \notin \{Z^n\}_{n=1}^M$. Finally, (a) holds since it holds for $\{Z^n\}_{n=1}^M$, and we know that $Z^n \subset Y \Rightarrow n \geq t$, and $Z^n \supset Y \Rightarrow n < t$. But $\{Z^n\}_{n=1}^M$ is the longest sequence in $\mathcal{Y}$, a contradiction.

Consequently, $Y \triangleright Z^1$, and (since we have already shown that $Z^n \supset Y \Rightarrow n < K$) we have $Y \not\subset Z^n$ for all n. Since (a) and (c) both hold for $\{Z^n\}_{n=1}^M$, it follows that $\{Y, \{Z^n\}_{n=1}^M\}$ satisfies (a) and (c). And since $\{Z^n\}_{n=1}^M$ satisfies (b) and $Y \notin \{Z^n\}_{n=1}^M$, we have that $\{Y, \{Z^n\}_{n=1}^M\}$ satisfies (b) as well. Then $\{Y, \{Z^n\}_{n=1}^M\} \in \mathcal{Y}$, a contradiction since $\{Z^n\}_{n=1}^M$ is the longest sequence in $\mathcal{Y}$.

Step 4: Construction of a strategically consistent profile. By Step 3 and since $\{Z^n\}_{n=1}^M \in \mathcal{Y}$, every element of $\mathcal{M}$ appears exactly once in $\{Z^n\}_{n=1}^M$. Hence, we can define a new strict total order $\succ$ on $\mathcal{M}$ as follows: $Z^n \succ Z^m \Leftrightarrow n < m$. Since $\{Z^n\}_{n=1}^M$ satisfies (c), this order refines $\supset$ on $\mathcal{M}$: If $Z^n \supset Z^m$, then $n < m$, and hence $Z^n \succ Z^m$. Let $\{C_i, \mu_i\}_{i \in I}$ be the profile of choice functions and beliefs constructed according to (3). By Lemma 7, $\{C_i, \mu_i\}_{i \in I}$ is strategically consistent.

Step 5: Common beliefs are rationalized by $\succ$: If $Y \succ Z$, then $\mu_i(Y \cup Z) = Y$ for each $i \in I$. Suppose $Y, Z \in \mathcal{M}$ are such that $Y \succ Z$. Since $\mathcal{M}$ is pairwise comparable, for all $S \subseteq Y \cup Z$, either $S \subseteq Y$ or $S \subseteq Z$. Then since $\succ$ refines $\supset$, for all $S \subseteq Y \cup Z$, either $S \preceq Y$ or $S \preceq Z$, and so by transitivity of $\succ$, $S \prec Y$. It follows that for each $i \in I$, $\mu_i(Y \cup Z) = \max_{\succ}\{Y' | Y' \subseteq Y \cup Z\} = Y$.

Step 6: Farsighted paths are $\succ$-increasing: If $Y \in \mathcal{M}$, and there is a farsighted path from Y to Z, then $Z \succ Y$. Suppose there is a farsighted path $\{Y^n\}_{n=0}^N$ from Y to Z. Since $\mu_i(Y^{n-1} \cup Y^n) = Y^n$ for each $i \in I$ and $n > 0$, by construction of μ, we have $Y^n \in \mathcal{M}$ for each $n > 0$. Since $Y^0 = Y \in \mathcal{M}$ by assumption, it follows from Step 5 that for each n, $Y^{n+1} \not\prec Y^n$. Then since $\succ$ is a strict total order on $\mathcal{M}$, we have $Y^{n+1} \succ Y^n$ for each $n < N$, and so by transitivity $Z \succ Y$.

Step 7: $\{C_i, \mu_i\}_{i \in I}$ satisfies forward induction. Suppose toward a contradiction that Z is a credible blocking proposal for $Y \in \mathcal{M}$, but $\mu_i(Z \cup Y) \neq Z$ for some $i \in I$. Then by

Step 5, $Z \not\succ Y$. By definition, $Z \in \mathcal{M}$; since, by construction, $\succ$ is a strict total order on $\mathcal{M}$, we have $Z \prec Y$. Moreover, since Z is a myopically credible blocking proposal for Y, and $\trianglerighteq$ is complete, $Z \triangleright Y$.

Since $Y, Z \in \mathcal{M}$, by Step 3, they must be elements of the sequence $\{Z^n\}_{n=1}^M$; label $Y = Z^y$ and $Z = Z^z$. By definition of $\succ$, since $Y \succ Z$, we must have $y < z$. Then since $Z^z \triangleright Z^y$, and $\{Z^n\}_{n=1}^M$ satisfies property (c), we must have $y < z-1$. Then $Y \succ Z^{z-1} \succ Z$, and so (by Step 5) there is a farsighted path from Z to Z^{z-1}, but (by Step 6) there is no farsighted path from Y to Z^{z-1}. Then Z is not a farsightedly credible blocking proposal for Y, a contradiction. □

Proof of Theorem 7 Suppose Y is stable given $\{\hat{C}_i\}_{i \in I}$. Then by definition, it is non-strategically individually rational: $\hat{C}_i(Y_i|Y_{-i}) = Y_i$ for each $i \in I$. Moreover, there is no $Y' \supset Y$ such that $\hat{C}_i(Y'_i|Y'_{-i}) = Y'_i$ for each $i \in I$: Suppose not, and there exists such a Y'. Then for all $i \in N(Y' \setminus Y)$, we have $Y'_i \setminus Y_i \subseteq Y'_i \subseteq \hat{C}_i(Y'_i|Y'_{-i})$, a contradiction since Y is stable (and therefore unblocked) given $\{\hat{C}_i\}_{i \in I}$. It follows from Theorem 5 that Y is stable for some strategically consistent assessment $\{C_i, \mu_i\}_{i \in I}$ satisfying weak forward induction.□