Preference Evolution under Partner Choice*

Ziwei Wang† Jiabin Wu‡

August 16, 2026

Abstract

We study preference evolution when agents choose whom to interact with. In the short run, subjective preferences influence both partner choice and strategic behavior, which affect their material payoffs. These payoffs, in turn, determine how preferences evolve in the long run. To model this "match-to-interact" process, we combine stable matching with correlated equilibrium, allowing matched agents to coordinate on self-enforcing agreements. We show that evolution under endogenous matching need not favor either purely selfish or purely efficient preferences. Instead, we identify two classes of preferences that combine these concerns nonlinearly and can prevail because they sustain efficient play while protecting agents from exploitation. Under incomplete information, however, only one of these classes retains its evolutionary advantage. This class combines efficiency with parochialism, a form of group-level selfishness that strengthens incentives for self-sorting.

Keywords: Preference Evolution, Stable Matching, Evolutionary Dominance, Incomplete Information, Morality, Parochialism
JEL Codes: C73, C78, Z10

*We thank Ingela Alger, Yi-Chun Chen, Jeffrey Ely, Wei He, Gaoji Hu, Luis Izquierdo, Segismundo Izquierdo, Sam Jindani, Jinwoo Kim, Ce Liu, Qingmin Liu, Zhuoran Lu, Weicheng Min, Jonathan Newton, Fanqi Shi, Xiang Sun, Yifei Sun, Ina Taneva, Qianfeng Tang, Yi Tong, Matthijs van Veelen, Jörgen Weibull, Yiqing Xing, Nate Yoder, Siyang Xiong, Hanzhe Zhang, Kun Zhang, and Jidong Zhou for constructive comments and suggestions. We also thank seminar and conference participants at Renmin University of China, Peking University, Capital University of Economics and Business, Shanghai Jiao Tong University, UC Riverside, NASM 2024 (Vanderbilt), the 7th World Congress of the Game Theory Society (PKU), the 2024 HUST Economic Theory Workshop, the LEG Online Seminar, CETC 2025 (Montréal), CoED 2025 (Essex), ESWC 2025 (Seoul), and the 2026 Summer Workshop on Microeconomic Theory (CUFE) for valuable discussions. Rongkai Liu provided helpful research assistance. Wang acknowledges financial support from the National Natural Science Foundation of China (No. 72403188). The usual disclaimer applies.

†Guanghua School of Management, Peking University. Email: zwang.econ@gmail.com.
‡Department of Economics, University of Oregon. Email: jwu5@uoregon.edu.

1 Introduction

Economic analysis typically treats decision makers' preferences as fixed and exogenous. Preferences, however, may themselves be the product of a lengthy evolutionary process. This raises a fundamental question: why do some preferences persist while others disappear? Güth and Yaari (1992) and Güth (1995) introduce the "indirect evolutionary approach," which provides a useful theoretical framework for understanding preference evolution: preferences shape behavior, behavior determines fitness, and fitness, in turn, governs the evolution of preferences.[1] A preference type is considered evolutionarily stable if, when predominant in a population, it can resist invasion by alternative types; see recent surveys by Alger and Weibull (2019) and Alger (2023).[2]

Most work in this tradition studies behavior in two-player games among individuals paired through some exogenous random matching process.[3] This focus, however, omits another important behavioral margin. Preferences may shape not only how individuals play after being matched, but also whom they choose to interact with and, hence, how matches form in the first place. When preferences determine both strategic behavior and partner choice, matching itself becomes endogenous in the evolutionary process.

This paper develops a model of preference evolution with endogenous matching that accommodates a rich space of preference types, any finite symmetric two-player game, and potentially incomplete information about other agents' preference types. We model the joint determination of matching and play in a continuum population by combining the noncooperative concept of correlated equilibrium and the cooperative notion of stable matching.[4] Previous studies by Jackson and Watts (2010), Xing (2020), and Garrido-Lucero and Laraki (2021) adopt a similar approach but rely on Nash equilibrium (or subgame-perfect Nash equilibrium) to discipline play within matches. The additional coordination permitted by correlated equilibrium is economically important when efficient interaction requires partners to take different roles. An agreement can then randomize over role assignments and thereby produce a more balanced division of expected payoff. Stability, in turn, captures the idea that unsatisfied individuals can communicate and jointly deviate by forming a new match and coordinating on an agreement that benefits both. A stable outcome can therefore be viewed as the reduced-form outcome of a frictionless process of partner formation.

We begin with a complete-information benchmark in which preference types are observable. To characterize stable outcomes in this environment, we propose a concept called *complete-information*

[1] Earlier contributions developing related ideas include Becker (1976), Hirshleifer (1977), Rubin and Paul (1979), and Frank (1987).

[2] See also Robson and Samuelson (2011) for a critical assessment of this approach.

[3] A few papers instead consider multilateral games (Lehmann et al., 2015; Alger and Weibull, 2016; Alger et al., 2020) or environments in which all individuals "play the field" (Lahkar, 2019; Bandhu and Lahkar, 2023). Matching remains exogenous in the former and plays no role in the latter.

[4] Formally, our framework is a large matching market with contracts. We draw on recent results in this literature to establish the existence of stable outcomes in our setting (Che et al., 2019; Greinecker and Kah, 2021; Jagadeesan and Vocke, 2024; Carmona and Laohakunakorn, 2024).

stability for a continuum population. Specifically, it requires that every matched pair of individuals plays a correlated equilibrium, termed *internal stability*, and that no pair of agents can coordinate on a correlated equilibrium that makes both strictly better off, termed *external stability*. Given a population composed of heterogeneous preference types, agents' preferences jointly shape both whom they match with and how they play within a match, and thereby the material payoffs they obtain in a stable outcome. We then evaluate the evolutionary performance of preference types by comparing their average material payoffs across stable outcomes.

We show that two natural benchmarks—selfish and efficient preferences—can each fail to prevail in evolution. Selfish preferences provide an inherent safeguard against exploitation because selfish agents always seek to maximize their own material payoffs. Yet they may sustain inefficient equilibrium play, leaving selfish agents at an evolutionary disadvantage.[5] Efficient preferences, by contrast, align agents' incentives with joint material payoff, which ensures that efficient play can be sustained in equilibrium. Their weakness is excessive benevolence: efficient agents may accept an unfavorable division of the joint material payoff and thus become vulnerable to exploitation.[6]

These weaknesses can be overcome by combining selfishness and efficiency in nonlinear ways. We identify two classes of preference types that do so and can thus prevail in evolution. The first consists of *morally constrained selfish* types, whose preferences rank the two objectives hierarchically. Agents with such preferences place first priority on efficient play with their partners and then, conditional on efficiency, maximize their own material payoffs. The efficiency objective disciplines behavior because, whenever agents of this type play inefficiently, two of them can profitably rematch and implement an efficient equilibrium. The selfish component, in turn, protects agents from unfavorable payoff divisions: agents who receive too little can rematch and use a symmetrized efficient equilibrium to divide the maximum joint material payoff equally. As a remarkable consequence of endogenous partner choice, agents with preferences in this class can achieve both efficiency in payoff realization and fairness in distribution.

The second class comprises *parochial efficient* types, whose preferences also embody a hierarchical ordering of objectives. Agents with such preferences exhibit a form of plasticity.[7] Under

[5]More generally, altruistic preferences that are not aligned with joint material payoff may suffer from the same weakness. This contrasts with the conventional view in theoretical biology, dating back to Hamilton (1964a,b), that assortative interaction—whether arising through kinship, reciprocity, or group selection—can sustain altruism. With endogenous partner choice, however, a competing type can itself generate positive assortative matching, turning inefficient play by altruistic agents into an evolutionary disadvantage.

[6]A preference for efficiency admits two interpretations. On the one hand, it can be viewed as a "disinterested" preference, whereby an individual is solely focused on the social objective of maximizing the total material payoff. This perspective is in line with *utilitarianism*: conditional on her partner's behavior, the individual regards an action as morally right when it maximizes the pair's welfare (Mill, 1863). On the other hand, the preference for efficiency can also be seen as a form of altruism, whereby an individual places equal importance on her own material payoff and that of her opponent.

[7]Plasticity refers to situations in which an individual's preferences depend not only on the action profile in the underlying game but also on her partner's preference type. Studies of preference evolution allowing for plasticity include Sethi and Somanathan (2001), Herold and Kuzmics (2009), Alger and Lehmann (2023), and Avataneo (2026).

this form of plasticity, which we call *parochialism*, agents derive utility exclusively from interactions with others of their own type.[8] Parochialism can therefore be understood as selfishness at the group level. Conditional on interacting with their own kind, these agents then seek to maximize joint material payoff. Types in this class can prevail in evolution for two reasons. First, parochialism fosters positive assortative matching, since in any stable outcome, agents with such preferences do not benefit from partnerships with agents of other types. Second, efficient play within the group guarantees them an average fitness at least as high as that of any competing type. This class therefore reveals a distinct route to evolutionary success—one based on selective association and efficient cooperation within the group.

Next, we turn to the case of incomplete information, under which individuals' preference types are their private information. Following the recent literature on stable matching with incomplete information (see, e.g., Liu et al., 2014; Liu, 2020; Chen and Hu, 2020, 2023; Liu, 2025; Wang, 2026; Catonini and Wang, 2026), we develop a notion of *incomplete-information stability* suited to our setting. Agents match on the basis of publicly observable *labels* and share correct beliefs about the type distribution associated with each label, but do not directly observe one another's types. As under complete information, internal stability requires self-enforcing play within existing matches, while external stability rules out mutually profitable rematching opportunities. The main conceptual challenge is to define blocking under incomplete information. Intuitively, a blocking proposal identifies a targeted type on each side and prescribes how both sides will play following the deviation. The targeted agents must find it optimal to follow the proposal and must strictly benefit from doing so. Because types are not directly observable, these conditions must continue to hold even if nontargeted types also join the deviation and behave differently from the targeted types. The resulting solution concept has two useful properties. First, it formally incorporates complete-information stability as a special case, thereby guaranteeing existence. Second, information is *endogenous* in a stable outcome, because the informational content conveyed by labels must itself be consistent with stability.

We show that morally constrained selfish types can be at an evolutionary disadvantage in a broad class of material games. Under complete information, agents of such a type can secure a high material payoff: if their current payoff is too low, they can identify one another, rematch, and coordinate on a fair and efficient equilibrium. Under incomplete information, however, the same deviation may no longer be attractive. Because a prospective partner's type is unobservable, a deviating agent must account for the possibility that nontargeted types also participate and respond adversely to the proposed play. The deviation may therefore fail to constitute an incomplete-information blocking pair in an unfavorable outcome, preventing morally constrained selfish agents from guaranteeing an average material payoff at least as high as that of a competing type.

[8] Parochialism describes a narrow concern for one's own group rather than the broader population. The term has been used in several related senses. Bernhard et al. (2006) define it as a preference for favoring members of one's own group, whereas Choi and Bowles (2007) define it as hostility toward other groups. Relatedly, the network literature commonly assumes *homophily*, a direct preference for associating with similar others; see Jackson (2014) for a survey.

In contrast, parochial efficient types continue to prevail in evolution even under incomplete information. Indeed, agents with such preferences do not face the same obstacle confronting morally constrained selfish types. Because they derive positive utility only from interacting with their own type and none from interacting with competing types, the behavior of nontargeted participants cannot eliminate their incentive to leave an unfavorable match and rematch with one another. We show that every incomplete-information stable outcome exhibits perfect assortative matching by label, and that within-label play is efficient for every label that may contain parochial efficient agents.[9] Together, these properties ensure that parochial efficient agents earn an average material payoff at least as high as that of any competing type, regardless of its population share.

The two papers most closely related to ours are Dekel et al. (2007) and Alger and Weibull (2013). Dekel et al. (2007) study preference evolution under uniform random matching and an exogenous degree of preference observability. They find that observability shapes which behaviors can survive evolutionary selection. When preference types are perfectly observable, only efficient outcomes can survive; when they are completely unobservable, by contrast, self-interested behavior becomes evolutionarily robust. They also consider intermediate degrees of observability and show that the necessity of efficient and selfish behaviors continues to hold in neighborhoods of the two polar cases. Our results under complete information similarly emphasize the force toward efficient play. Our main departure, however, arises under incomplete information. With stable partner choice, the information conveyed by observable labels is itself endogenous. We show that parochial efficient types can overcome the informational friction involved in forming blocking pairs, sustain efficient play, and thus prevail in evolution even when types are unobservable.

Alger and Weibull (2013) also study preference evolution under incomplete information but impose an exogenous matching process with a given degree of assortativity. Their central result establishes a sharp link between matching and morality: *homo moralis*, whose utility is a convex combination of own material payoff and a Kantian moral objective, is evolutionarily successful when the weight on the Kantian term equals the degree of assortativity. Our contribution is to derive assortativity endogenously from stable partner choice. Our results reinforce their broader insight that evolution can favor preferences combining self-interest with a moral concern, but the resulting preferences differ in two crucial respects. First, the moral objective is defined differently. Whereas the Kantian criterion in Alger and Weibull (2013) evaluates an action by the payoff it would yield if the partner chose the same action, moral play in our model maximizes joint material payoff. It may therefore require the partners to take different roles and actions, which is made possible by coordination within a match. Second, our types combine self-interest and the moral concern in a

[9] Although our analysis characterizes stable outcomes rather than the adjustment process leading to them, the results admit a heuristic interpretation of partial unraveling. Starting from coarse labels, parochial efficient agents can form a blocking pair whenever the matching is not assortative or their play is inefficient. If participation in and behavior following deviations are publicly observed and informative, rematching can promote assortative matching directly and, at the same time, generate information about agents' underlying types that can be incorporated into finer labels.

lexicographic manner: the lower-priority concern matters only when the higher-priority concern is satisfied.[10]

1.1 Related Literature

Frank (1987) studies preference evolution with endogenous matching under incomplete information. His model imposes an exogenous information structure that induces positive assortative matching, after which matched agents interact only once. Frank shows that commitment power and the ability to identify committed partners are key drivers of evolutionary success. Our results operate through a related mechanism. Preferences that induce efficient play serve as a form of conditional commitment, while parochialism combined with stable partner choice allows type identification and assortative matching to arise endogenously. Subsequent studies by Nowak et al. (2000), Akdeniz and van Veelen (2021), and Akdeniz and van Veelen (2023) build on this insight, showing how evolution can favor commitment to behavior that would otherwise be irrational.

In a random-matching environment under incomplete information, Güth and Kliemt (1998) demonstrate that conditional cooperators cannot survive when preference types are unobservable, because they cannot tailor their behavior to their opponents' types. Subsequent contributions include Ely and Yilankaya (2001), Ok and Vega-Redondo (2001), and Dekel et al. (2007). More recently, Avataneo et al. (2025) show that the five principles of Moral Foundations Theory generically span the set of evolutionarily stable preferences under random matching and observable types. Both Herold and Kuzmics (2009) and Avataneo (2026) allow plastic preferences which depend on the opponent's type, as we do in this paper. Herold and Kuzmics (2009) show that discriminating types are evolutionarily stable under (almost) complete information, while Avataneo (2026) finds that universalism is favored when repeated interaction with the same partner is unlikely and particularism when it is likely. We also recognize the importance of plasticity but show that it plays a different role under endogenous matching: it shapes not only how agents behave toward different types, but also whom they choose to interact with and, consequently, the resulting matching patterns.

Alger and Weibull (2013) study preference evolution under incomplete information and exogenous matching with a given degree of assortativity. Newton (2017) extends Alger and Weibull (2013) by allowing the degree of matching assortativity to evolve. He demonstrates that Kantian preferences can prevail when coupled with parochialism defined through the matching rule. Wu (2020) studies observable labels that are exogenously correlated with otherwise unobservable preference types and assumes homophilic matching on those labels. In contrast, our model embeds parochialism directly in preferences and endogenizes, through stable partner choice, both the information conveyed by labels and the resulting degree of assortativity. Our approach is also related to models of endogenous interaction structures and dynamic partner choice (see Jackson and Watts, 2002; Fujiwara-Greve

[10] Any convex combination of selfishness and material efficiency does not prevail in our model. For every such preference type, there exists a material game in which it is evolutionarily disadvantaged.

and Okuno-Fujiwara, 2009; Izquierdo et al., 2010, 2014, 2021; Graser et al., 2024, among others).

2 Population, Material Game, and Preference Types

Consider a continuum population of unit mass. Agents are matched in pairs and interact through a symmetric game $\mathcal{G}$. The game has a finite common action set A. If an agent plays action $a \in A$ while her partner plays action $a' \in A$, she receives a **material payoff** (or **fitness**) $\pi(a, a')$, where $\pi : A^2 \to \mathbb{R}$. Because A^2 is finite and every agent participates in exactly one match, adding the same constant to all material payoffs does not affect fitness comparisons. We therefore normalize $\pi(a, a') > 0$ for all $(a, a') \in A^2$.[11]

Write Θ for the set of **preference types**. Formally, a preference environment is a pair (Θ, u), where

$$u : A^2 \times \Theta \times \Theta \to \mathbb{R}.$$

For $\theta, t \in \Theta$, write $u_\theta(a, a', t) = u(a, a', \theta, t)$ for the utility of a type-θ agent who plays a against a type-t partner who plays a'. We interpret Θ as the set of preference types that may arise through mutation and assume that it includes all types constructed in the following analysis.[12] This specification extends the standard outcome-based formulation in the literature on preference evolution by allowing utility to depend on the matched partner's preference type. We call type θ **plastic** if its utility depends on its partner's type; that is, if $u_\theta(a, a', t) \neq u_\theta(a, a', t')$ for some $t, t' \in \Theta$ and $(a, a') \in A^2$.

We make two additional remarks. First, we focus on material games with a finite common action set A, as in Dekel et al. (2007). Our analysis extends to games with a general topological strategy space under standard regularity conditions, but the central insights of the paper remain unchanged. Second, we impose no restriction linking the material payoff function π to the utility function u_θ. Thus, agents may maximize objectives other than their own material fitness. The following example illustrates the flexibility allowed by this formulation.

Example 1. The model accommodates a wide range of preferences.

(i) A **selfish** type cares only about her own material payoff: $u_\theta(a, a', t) = \pi(a, a')$.

(ii) An **efficient** type cares about the total material payoff generated within the matched pair: $u_\theta(a, a', t) = \pi(a, a') + \pi(a', a)$.

(iii) A **behavioral** type may have an objective unrelated to material payoffs. For example, for some fixed action $\bar{a} \in A$, $u_\theta(a, a', t) = \mathbf{1}_{\{a=\bar{a}\}}$.

[11] We do not allow agents to opt out of playing the material game. Nevertheless, our analysis does extend to a setting in which agents can remain unmatched, provided that the outside option yields a material payoff strictly below the lowest payoff attainable from any match—for example, zero under our normalization.

[12] This reduced-form specification follows Herold and Kuzmics (2009) (see also Avataneo, 2026). Gul and Pesendorfer (2016) provide a foundation for such interdependent preferences by constructing a canonical type space from characteristics and hierarchies of preference responses.

(iv) A **plastic** type conditions its concern for the partner's payoff on the partner's type. For example, $u_\theta(a, a', t) = \pi(a, a') + \pi(a', a) \cdot \mathbf{1}_{\{t=\theta\}}$. ◇

When two agents interact, they may coordinate their behavior through an **agreement** $\alpha \in \Delta(A^2)$, defined as a probability distribution over action pairs. Let $\mathcal{A} = \Delta(A^2)$ denote the set of all possible agreements. We impose no restrictions on α, so agreements may involve arbitrary correlation between agents' actions. This formulation is deliberately general. An agreement may be deterministic, placing probability one on a single action pair (a, a'). It may also specify an independent joint distribution, corresponding to the agents' use of mixed strategies. More generally, agents may coordinate their actions through a randomizing device that generates public or private action recommendations.

Randomizing devices are common in real-world interactions and provide a natural way for agents to coordinate behavior. This is particularly relevant in an evolutionary setting, where agents interact over a long horizon and may learn to utilize such devices. Consider the familiar Battle of the Sexes, in which agents benefit from coordination but disagree about which activity to coordinate on.[13] A public coin flip can eliminate miscoordination while treating the agents symmetrically ex ante. It can also be viewed as a reduced-form representation of a long-run arrangement such as turn-taking.

For our main analysis, we consider a population with two distinct preference types $\theta, \tau \in \Theta$.[14] A proportion $1 - \varepsilon$ of the agents are of type θ, and the remaining proportion ε are of type τ, where $\varepsilon \in (0, 1)$. We refer to the tuple $(\theta, \tau, \varepsilon)$ as a **population state**.

3 Preference Evolution with Complete Information

In this section, we assume that each agent observes the preference types of all other agents. Hence, when two agents are matched, they play $\mathcal{G}$ with complete information.

Fix a population state $(\theta, \tau, \varepsilon)$. For each type $t \in \{\theta, \tau\}$, we let $\mu_t \in \Delta(\{\theta, \tau\})$ be a probability distribution over types in the population that describes how type-t agents are matched. A **matching profile** is a vector $\mu = (\mu_\theta, \mu_\tau)$ that satisfies the following consistency condition:

$$(1 - \varepsilon)\mu_\theta[\tau] = \varepsilon\mu_\tau[\theta].$$

This condition requires that the total mass of type-θ agents matched with type-τ agents is equal to that of type-τ agents matched with type-θ agents.[15]

For any $t, t' \in \{\theta, \tau\}$, let $\gamma_{t,t'} \in \Delta(\mathcal{A})$ describe the conditional distribution of agreements reached across matches between type-t and type-t' agents. Let $\rho : \mathcal{A} \to \mathcal{A}$ be a mapping that

[13] Formally, the Battle of the Sexes can be represented as a symmetric two-player game with a common action set by defining actions in player-relative terms: each player chooses either her own preferred activity or her partner's preferred activity.

[14] In Section 5, we discuss how our results extend to the more general case of polymorphic populations.

[15] Although we take a distributional approach in defining the matching profile, there is no randomness in how agents are matched. Given that the population consists of finitely many types, we can explicitly describe the matching pattern through a deterministic mapping that generates μ.

reverses the roles of the two agents; that is, for any $\alpha \in \mathcal{A}$,

$$\rho(\alpha)[a, a'] = \alpha[a', a] \text{ for all } (a, a') \in A^2.$$

An **agreement profile** $\gamma = (\gamma_{t,t'})$ is a vector of conditional distributions of agreements that satisfies the following exchangeability condition: $\gamma_{t,t'}[B] = \gamma_{t',t}[\rho(B)]$ for any measurable set $B \subseteq \mathcal{A}$.

We call the combination of a matching profile and an agreement profile (μ, γ) an **outcome**.

3.1 Stable Outcome

Given a population state $(\theta, \tau, \varepsilon)$, our next goal is to identify the outcomes (μ, γ) that can be deemed *stable*. The requirement of stability has two layers. First, holding the matching profile μ fixed, agents do not want to change their behavior as specified by γ. In other words, every agreement should induce equilibrium play. Second, given the utilities agents derive in an outcome, there should not exist agents who want to form a pairwise deviation and mutually benefit from rematching. To formalize this idea, we extend the stability notion of Garrido-Lucero and Laraki (2021) to a continuous population and from Nash equilibrium to correlated equilibrium; see also Jackson and Watts (2010).

We now formally define these two layers of stability. For each agreement $\alpha \in \mathcal{A}$ between t and t' and a function $g : A \to A$, let

$$\mathbb{E}_\alpha[u_t(g(a), a', t')] = \sum_{(a,a') \in A^2} u_t(g(a), a', t')\alpha[a, a']$$

denote the expected utility type-t agent obtains by playing $g(a)$ whenever the agreement recommends action a. Let $f : A \to A$ be the strategy of *following* the recommendation so that $f(a) = a$ for all $a \in A$. For $t, t' \in \{\theta, \tau\}$, let $CE_{t,t'}$ denote the set of correlated equilibria between type-t and type-t' agents. That is, $\alpha \in CE_{t,t'}$ if and only if

$$f \in \arg\max_g \mathbb{E}_\alpha[u_t(g(a), a', t')] \quad \text{and} \quad f \in \arg\max_g \mathbb{E}_{\rho(\alpha)}[u_{t'}(g(a), a', t)].$$

As discussed in Section 2, an agreement can be interpreted as a correlation device that draws an action pair and privately recommends one action to each agent. A correlated equilibrium requires the agreement to be self-enforcing: conditional on her own recommendation, each agent finds it optimal to follow it when her partner does the same (Aumann, 1974, 1987). This solution concept is natural in our setting. Matched agents can coordinate before play but cannot commit to their actions. Correlated equilibrium captures precisely this combination: it permits coordination while requiring its recommendations to be voluntarily followed.[16]

[16]The direct-recommendation formulation is without loss of generality. More generally, agents could agree ex ante

An agreement profile is a correlated equilibrium profile if almost every matched pair plays a correlated equilibrium.

Definition 1. An agreement profile γ is a **correlated equilibrium profile** with respect to μ if for every $t, t' \in \{\theta, \tau\}$ such that $\mu_t[t'] > 0$, $\gamma_{t,t'}[CE_{t,t'}] = 1$.

A block specifies two agents who are present in matches with positive mass, together with an agreement for their rematch. The block is viable under complete information if the proposed agreement is self-enforcing for the two deviating types and makes both of them strictly better off compared to the status quo. For $t \in \{\theta, \tau\}$, call a pair $(\bar{t}, \bar{\alpha}) \in \{\theta, \tau\} \times \mathcal{A}$ a **current position** of type t under (μ, γ) if $\mu_t[\bar{t}] > 0$ and $\bar{\alpha} \in \text{supp}(\gamma_{t,\bar{t}})$.

Definition 2. Fix an outcome (μ, γ). A **complete-information blocking pair** exists if there are types $t, t' \in \{\theta, \tau\}$; current positions $(\bar{t}, \bar{\alpha})$ of type t and $(\bar{t}', \bar{\alpha}')$ of type t'; and an agreement $\hat{\alpha} \in \mathcal{A}$ such that

(i) $f \in \arg\max_g \mathbb{E}_{\hat{\alpha}}[u_t(g(a), a', t')]$ and $f \in \arg\max_g \mathbb{E}_{\rho(\hat{\alpha})}[u_{t'}(g(a), a', t)]$;

(ii) $\mathbb{E}_{\hat{\alpha}}[u_t(a, a', t')] > \mathbb{E}_{\bar{\alpha}}[u_t(a, a', \bar{t})]$ and $\mathbb{E}_{\rho(\hat{\alpha})}[u_{t'}(a, a', t)] > \mathbb{E}_{\bar{\alpha}'}[u_{t'}(a, a', \bar{t}')]$.

This notion of blocking for aggregate matching is analogous to the one in Echenique et al. (2013), which is a natural generalization of the blocking concept proposed by Gale and Shapley (1962) to continuous populations. For a continuous population, deviating agents in a blocking pair must have positive mass; otherwise, they would be unable to locate one another. Moreover, the proposed agreement must satisfy two requirements. First, it must be self-enforcing: the deviating agents coordinate on a correlated equilibrium, so following the action recommendation is optimal for each side. Second, it must be profitable: both agents strictly prefer the proposed rematch to their current positions.[17]

For an outcome to be stable, almost every matched pair should play a correlated equilibrium, and no blocking pair should be able to upset the outcome. This leads to the following definition.

Definition 3. An outcome (μ, γ) is **complete-information stable** if it satisfies:

(i) (Internal stability) γ is a correlated equilibrium profile with respect to μ;

(ii) (External stability) There is no complete-information blocking pair.

on an arbitrary payoff-irrelevant signal structure and play a Bayes-Nash equilibrium conditional on their signals. The resulting distribution of action pairs is a correlated equilibrium (Aumann, 1987).

[17] Our stability notion requires both parties to have strict incentives to rematch. This requirement is standard in the literature on matching games without transfers, as it is robust to small payoff perturbations or infinitesimal rematching costs. A weaker notion of blocking that imposes a weak incentive on one side can sharpen the set of stable outcomes but is known to generate unappealing non-existence issues. We use the strict-gain formulation as a conservative and robust test for whether a proposed rematch should count as a blocking pair.

Complete-information stability describes outcomes that, once reached, leave no room for further strategic or coalitional adjustment. In the spirit of pairwise stability in matching games, agents evaluate deviations myopically: they compare the one-shot payoff from a proposed rematch with their current payoff, without considering how the deviation may affect future opportunities.

Since our model has no predetermined sides, the matching problem resembles a "roommate problem," where stable matchings need not exist in finite economies (Gale and Shapley, 1962). In our continuum environment, however, stable outcomes always exist—a result we prove in Proposition 7 for the more general case of polymorphic populations. The next example illustrates Definition 3.

Example 2. Consider the following underlying material game and a population state $(\theta, \tau, \varepsilon)$.

	x	y
x	$1, 1$	$4, 2$
y	$2, 4$	$1, 1$

Assume that type θ is selfish, i.e., $u_\theta(a, a', t) = \pi(a, a')$, whereas the preference type τ is behavioral, with utility function $u_\tau(a, a', t) = \mathbf{1}_{\{a=x\}}$. Since action x is dominant for type-τ agents, the unique correlated equilibrium $\hat{\alpha}$ between two type-τ agents satisfies $\hat{\alpha}[x, x] = 1$.

We now argue that any complete-information stable outcome (μ, γ) must satisfy $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$; $\gamma_{\theta,\theta}[\tilde{\alpha}] = 1$, where $\tilde{\alpha}[x, y] = \tilde{\alpha}[y, x] = 1/2$; $\gamma_{\tau,\tau}[\hat{\alpha}] = 1$.[18] That is, matching is perfectly assortative. Every matched pair of type-θ agents plays (x, y) and (y, x) with equal probability, which can be implemented as a correlated equilibrium with public randomization. Each matched pair of type-τ agents plays (x, x).

To see this, suppose the contrary. There are two cases to consider:

- Suppose $\mu_\theta[\tau] > 0$. Internal stability then implies that the unique agreement $\alpha \in \text{supp}(\gamma_{\theta,\tau})$ satisfies $\alpha[y, x] = 1$, so type-θ agents in cross-type matches obtain utility 2. But any two such type-θ agents can block by rematching with each other and playing the correlated equilibrium $\tilde{\alpha}$, which gives each of them expected utility 3. This contradicts external stability.
- Suppose $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$, but $\text{supp}(\gamma_{\theta,\theta}) \neq \{\tilde{\alpha}\}$. Then a positive mass of type-θ agents obtain expected utility strictly below 3: in any correlated equilibrium between two type-θ agents, the *lower* of the two agents' expected utilities is at most 3, and this bound is attained only at $\tilde{\alpha}$. These agents can therefore block by rematching with one another and playing $\tilde{\alpha}$, again violating external stability. ◇

Example 2 illustrates an *equilibrium-selection* effect of stability. There are many correlated equilibria between two type-θ agents, but only $\tilde{\alpha}$ can arise in a complete-information stable outcome.

[18] The complete-information stable outcome is unique in a generic sense: $\gamma_{\theta,\tau}$ may be specified arbitrarily for a measure-zero set of cross-type matches and has no bearing on complete-information stability.

External stability rules out the others: if a correlated equilibrium leaves some type-θ agents below an equal split of the efficient total payoff, those agents can profitably rematch and play $\tilde{\alpha}$.[19]

3.2 Evolutionary Dominance

Following the indirect evolutionary approach, we assume that, at the behavioral level, the population reaches a stable outcome before evolutionary forces appreciably alter the distribution of types. At the preference level, the type distribution evolves according to the material payoffs earned by different types in stable outcomes. Importantly, stable outcomes are determined by agents' subjective preferences, whereas evolutionary selection is governed by material success.

Given a complete-information stable outcome (μ, γ) in population state $(\theta, \tau, \varepsilon)$, the average material payoffs for type-θ and type-τ agents are given by

$$G_\theta(\mu,\gamma) = \sum_{t\in\{\theta,\tau\}} \left[\int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a,a')]\, d\gamma_{\theta,t}[\alpha]\right] \mu_\theta[t],$$
$$G_\tau(\mu,\gamma) = \sum_{t\in\{\theta,\tau\}} \left[\int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a,a')]\, d\gamma_{\tau,t}[\alpha]\right] \mu_\tau[t].$$

We now define a static notion that captures the evolutionary selection process, called evolutionary dominance, as follows.

Definition 4. Fix a material game $\mathcal{G}$.

(i) A preference type $\theta \in \Theta$ is **evolutionarily dominant** if for every $\tau \in \Theta$ and every $\varepsilon \in (0,1)$, every complete-information stable outcome (μ, γ) in the population state $(\theta, \tau, \varepsilon)$ satisfies $G_\theta(\mu,\gamma) \geq G_\tau(\mu,\gamma)$.

(ii) A preference type $\theta \in \Theta$ is **evolutionarily dominated** if there exists another type $\tau \in \Theta$ such that for every $\varepsilon \in (0,1)$, every complete-information stable outcome (μ, γ) in the population state $(\theta, \tau, \varepsilon)$ satisfies $G_\theta(\mu,\gamma) \leq G_\tau(\mu,\gamma)$, with strict inequality for some complete-information stable outcome.

Remark 1. Our definition of evolutionary dominance is closely related to the notions of evolutionary stability in Dekel et al. (2007) and Alger and Weibull (2013), but it differs in two respects. First, those papers fix an exogenous matching technology and compare average material payoffs across equilibrium play; in contrast, we endogenize matching and compare payoffs across stable outcomes. Second, their notions are defined in a local sense, while ours is global: the payoff comparison should hold for every population share ε of the competing type τ. Hence, an evolutionarily dominant type θ must both resist invasion when it is the incumbent type (i.e., ε is close to 0) and have the ability to invade when it is the mutant type (i.e., ε is close to 1).

[19] This effect is similar in spirit to that analyzed by Jackson and Watts (2010). In both settings, stability restricts the outcome and thereby refines the set of equilibria that can arise.

Two natural benchmarks—the selfish and efficient types—capture two forces central to evolutionary success, but neither is robust on its own. Selfish preferences protect agents from exploitation but may sustain inefficient play, whereas efficient preferences promote material efficiency but may expose agents to unfavorable payoff divisions. We defer a formal analysis of these limitations to Section 3.3 and first introduce two new classes of preference types that combine selfishness and efficiency nonlinearly. One subjects selfish behavior to an efficiency constraint; the other makes efficiency concerns conditional on interacting with one's own type. Our main result under complete information is that types in both classes are evolutionarily dominant.

Let m denote the maximum joint material payoff generated by a pure action pair; that is

$$m = \max_{(a,a')\in A^2} \pi(a,a') + \pi(a',a).$$

Moreover, let $\mathcal{E} \subseteq A^2$ be the set of action pairs that attain this joint payoff:

$$\mathcal{E} = \left\{(a,a') \in A^2 : \pi(a,a') + \pi(a',a) = m\right\}.$$

An agreement α is **efficient** if it assigns probability one to $\mathcal{E}$, i.e., $\alpha[\mathcal{E}] = 1$; or equivalently, $\mathbb{E}_\alpha[\pi(a,a') + \pi(a',a)] = m$.[20] An agreement that is not efficient is **inefficient**.

Definition 5. A preference type θ is **morally constrained selfish** if, up to a positive affine transformation, its utility function can be written as

$$u_\theta(a,a',t) = \mathbf{1}_{\{(a,a')\in\mathcal{E}\}} \cdot \pi(a,a'). \tag{1}$$

Because positive affine transformations preserve expected-utility comparisons, we henceforth adopt the normalization in (1) without loss of generality. We use the same convention for all preference classes defined below.

Recall that $\mathcal{E}$ is the set of efficient action pairs. A morally constrained selfish type then embodies a hierarchical ordering of objectives: agents with this preference type first seek efficient interaction with their partner and then, among efficient outcomes, choose the action that maximizes their own material payoff. We have the following positive result:

Proposition 1. *For any material game $\mathcal{G}$, a morally constrained selfish type is evolutionarily dominant.*

The intuition for Proposition 1 is as follows. Let θ denote a morally constrained selfish type. The moral constraint ensures that there exists an efficient agreement $\alpha \in CE_{\theta,\theta}$, so two type-$\theta$ agents

[20] In previous literature under exogenous matching, since agents of the same type are assumed to play the same action in equilibrium, the consideration of efficiency is restricted to symmetric action profiles. See, for example, Dekel et al. (2007).

can sustain efficient play (Lemma 1). The selfish component and external stability then guarantee that almost every type-θ agent obtains expected material payoff at least $m/2$. For if not, two such agents can form a blocking pair by playing $\alpha/2 + \rho(\alpha)/2$: such an agreement delivers $m/2$ and also belongs to $CE_{\theta,\theta}$ by symmetry and convexity of the set of correlated equilibria. Because no matched pair can generate total material payoff above m, no other type can earn an average material payoff above $m/2$ when type θ earns no less than $m/2$. The morally constrained selfish type therefore performs at least as well as any competing type in every stable outcome.

Definition 6. A preference type θ is **parochial efficient** if, up to a positive affine transformation, its utility function can be written as

$$u_\theta(a, a', t) = [\pi(a, a') + \pi(a', a)] \cdot \mathbf{1}_{\{t=\theta\}}. \tag{2}$$

The utility function (2) of a parochial efficient type also has a hierarchical structure. At the first level, such agents derive utility only from interacting with partners who share their preference type, a form of plasticity we call *parochialism*. Their first-order objective is therefore to sort with their own type. We interpret this parochial concern as selfishness at the group level. Conditional on same-type matching, agents then seek to maximize joint material payoff.

Newton (2017) also considers parochialism in preference evolution. He defines it at a matching level: parochial agents are assumed to match only with one another. In contrast, we define parochialism at the preference level. Whether parochial agents actually sort with their own type is determined endogenously by stable partner choice.

We show that a parochial efficient type can also prevail in the long run.

Proposition 2. *For any material game $\mathcal{G}$, a parochial efficient type is evolutionarily dominant.*

The mechanism behind Proposition 2 differs from that behind Proposition 1. The evolutionary dominance of a parochial efficient type rests on two behavioral implications: assortative matching and efficient play. The parochial component gives agents an incentive to match with their own type, while the efficiency component leads them to maximize joint material payoff within same-type matches. Assortative matching protects these agents from exploitation by other types. Thus, the vulnerability of a purely efficient type (see Example 4 below) is neutralized: when efficient play requires one agent to accept a lower material payoff, the corresponding gain accrues to another agent with the same preferences and therefore benefits the same evolutionary lineage.

While the preceding results identify nonlinear combinations of efficiency and selfishness that are sufficient for evolutionary dominance, they do not exhaust the preference types favored by evolutionary pressure. Nevertheless, the following result provides a necessary condition for a preference type to prevail.

Proposition 3. *Fix a material game $\mathcal{G}$ and a preference type θ. If there exists an agreement*

$$\hat{\alpha} \in \arg\max_{\alpha \in CE_{\theta,\theta}} \mathbb{E}_{\alpha}[u_{\theta}(a, a', \theta) + u_{\theta}(a', a, \theta)]$$

that is inefficient, then θ is evolutionarily dominated.

To understand this result, first observe that any correlated equilibrium has a fair counterpart with the same joint expected utility: assigning its two roles symmetrically equalizes the agents' expected utilities while preserving their incentives to follow its recommendations. Moreover, recall that stable partner choice has an equilibrium-selection effect. In same-type matches, only correlated equilibria that are fair and attain the highest joint expected utility can be sustained. Otherwise, two agents in disadvantageous positions can form a blocking pair. This selection effect, however, operates according to agents' preferences rather than material payoffs. If an inefficient correlated equilibrium attains the highest joint expected utility, its fair version remains inefficient. Materially inefficient play can therefore persist in same-type matches. Consequently, such a preference type is evolutionarily dominated by a competing type that can force assortative matching and efficient play.

An implication of Proposition 3 is that, if a population consists of a single type that is evolutionarily dominant, then almost every pairwise interaction must be efficient in a stable outcome (see a generalization of this insight in Proposition 8(ii)). This conclusion is consistent with that of Dekel et al. (2007), obtained under complete information in a random-matching environment.

3.3 Pure Selfishness and Efficiency

We now return to two familiar benchmarks: the selfish and efficient types. Up to positive affine transformations, their utility functions are, respectively, $u_{\theta}(a, a', t) = \pi(a, a')$ and $u_{\theta}(a, a', t) = \pi(a, a') + \pi(a', a)$. These types isolate two key forces emphasized in the literature on preference evolution: resistance to exploitation and the ability to sustain efficient play. The following proposition shows, however, that neither force alone guarantees evolutionary dominance without additional restrictions on the material game.

We call an agreement a correlated equilibrium of the material game if it is a correlated equilibrium when both agents' utilities coincide with their material payoffs.

Proposition 4. *Either a selfish type or an efficient type may be evolutionarily dominated for some material game $\mathcal{G}$.*

(i) If there exists an efficient correlated equilibrium of the material game, then a selfish type is evolutionarily dominant; otherwise, a selfish type is evolutionarily dominated.

(ii) If $\pi(a, a') = \pi(a', a)$ for every $(a, a') \in \mathcal{E}$, then an efficient type is evolutionarily dominant; otherwise, an efficient type is evolutionarily dominated.

For a selfish type, the evolutionary weakness lies in its failure to promote material efficiency: selfish agents may become trapped in an inefficient correlated equilibrium. The following example illustrates this possibility.

Example 3. Suppose that the underlying material game is given as follows:

	x	y
x	$1, 1$	$2, 8$
y	$8, 2$	$3, 3$

Let θ denote a selfish type and τ a parochial efficient type. Fix any $\varepsilon \in (0, 1)$. By Proposition 2, $G_\tau(\mu, \gamma) \geq G_\theta(\mu, \gamma)$ in every complete-information stable outcome. It therefore remains to construct a stable outcome in which the inequality is strict.

Consider the outcome (μ, γ) with perfectly assortative matching, $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$. In type-$\theta$ matches, let $\gamma_{\theta,\theta}[\hat{\alpha}] = 1$, where $\hat{\alpha}[y, y] = 1$. In type-τ matches, let $\gamma_{\tau,\tau}[\tilde{\alpha}] = 1$, where $\tilde{\alpha}[x, y] = \tilde{\alpha}[y, x] = 1/2$.

Internal stability follows because y is strictly dominant for selfish agents, while following $\tilde{\alpha}$ maximizes the joint material payoff of parochial efficient agents matched with their own type. For external stability, every type-τ agent obtains utility 10, the highest feasible utility, so no blocking pair can involve type τ. Moreover, $\hat{\alpha}$ is the unique correlated equilibrium between two type-θ agents and gives each of them the same material payoff as in the status quo. Hence, no two type-θ agents can form a blocking pair. The outcome is therefore complete-information stable.

In this outcome,

$$G_\tau(\mu, \gamma) = 5 > 3 = G_\theta(\mu, \gamma).$$

Because ε was arbitrary, the selfish type θ is evolutionarily dominated by the parochial efficient type τ. ◇

An efficient type faces the opposite problem: its benevolent concern for joint material payoff leaves it indifferent to how that payoff is divided, which makes it vulnerable to exploitation. To see this, consider the following example.

Example 4. Suppose that the material game is the same as in Example 3. Let θ denote an efficient type and τ a morally constrained selfish type. By Proposition 1, $G_\tau(\mu, \gamma) \geq G_\theta(\mu, \gamma)$ in every complete-information stable outcome. It therefore remains to construct a stable outcome in which the inequality is strict.

Fix any $\varepsilon \in (0, 1)$ and choose $c \in (0, \min\{\varepsilon, 1 - \varepsilon\})$. Consider a matching profile satisfying $(1-\varepsilon)\mu_\theta[\tau] = \varepsilon\mu_\tau[\theta] = c$, with the remaining probability assigned to same-type matches. In cross-type matches, let $\gamma_{\theta,\tau}[\hat{\alpha}] = 1$, where $\hat{\alpha}[x, y] = 1$. In same-type matches, let $\gamma_{\theta,\theta}[\tilde{\alpha}] = \gamma_{\tau,\tau}[\tilde{\alpha}] = 1$, where $\tilde{\alpha}[x, y] = \tilde{\alpha}[y, x] = 1/2$.

Internal stability follows because every prescribed action pair is efficient. Type-θ agents therefore obtain their highest feasible utility, while a type-τ agent can deviate only by making the action pair inefficient, which gives her utility zero.

For external stability, every type-θ agent already obtains the highest feasible utility, so no blocking pair can involve type θ. Type-τ agents obtain expected utility 5 in same-type matches and utility 8 in cross-type matches. Because no agreement between two type-τ agents can generate joint expected utility greater than 10, no two such agents can both strictly benefit from rematching. The outcome is therefore complete-information stable.

Finally,

$$G_\tau(\mu,\gamma) = 5\left(1-\frac{c}{\varepsilon}\right) + 8\frac{c}{\varepsilon} > 5 > 5\left(1-\frac{c}{1-\varepsilon}\right) + 2\frac{c}{1-\varepsilon} = G_\theta(\mu,\gamma).$$

Because ε was arbitrary, the efficient type θ is evolutionarily dominated by the morally constrained selfish type τ. ◊

Remark 2. One may wonder whether any convex combination of selfish and efficient preferences can prevail across material games. The answer is no. Indeed, any nontrivial convex combination yields a preference type with utility function $u_\theta(a,a',t) = \pi(a,a') + \beta\pi(a',a)$ for some $\beta \in (0,1)$. For every such β, one can construct a material game in which the corresponding type satisfies the condition in Proposition 3 and is therefore evolutionarily dominated. Thus, preferences of this form face the same problem as selfish types do and may fail to sustain efficient play in equilibrium.

4 Preference Evolution with Incomplete Information

In this section, we turn to incomplete information. Suppose that each agent knows her own preference type but may not observe the preference types of other agents in the population.[21] For a fixed population state $(\theta,\tau,\varepsilon)$, we therefore generalize the notion of a matching profile to accommodate incomplete information.

A **matching profile** (with incomplete information) is a tuple (Λ, p, q, μ). The first component Λ is a finite set of **labels** that are publicly observable.[22] There is a probability distribution with full support $p \in \Delta(\Lambda)$ over the set of labels. For each $\lambda \in \Lambda$, let $q_\lambda \in \Delta(\{\theta,\tau\})$ denote the distribution of preference types conditional on label λ. We assume $q_\lambda \neq q_{\lambda'}$ whenever $\lambda \neq \lambda'$, reflecting the

[21] An agent may care about certain characteristics of a potential partner, which, for simplicity, we assume uniquely determine the partner's preferences. When these characteristics are readily observable (e.g., physical appearance), a model with complete information suffices. When they are not directly observable (e.g., empathy or a sense of responsibility), however, incomplete information must be taken into account. A richer formulation could treat characteristics and preference types as separate primitives (see, for example, Gul and Pesendorfer, 2016), but we do not pursue that extension here.

[22] Labels are purely informational: they convey information about preference types but do not directly enter agents' utilities.

fact that different labels convey distinct information, and write $q = (q_\lambda)_{\lambda\in\Lambda}$. The pair (p, q) should satisfy the following marginal condition:

$$\sum_{\lambda\in\Lambda} p[\lambda]q_\lambda[\theta] = 1 - \varepsilon.$$

In words, the masses of type-θ agents with different labels should sum up to their total mass in the population. For brevity, we shall refer to a type-θ agent with label λ as a type-θ_λ agent. As in the case of complete information, for each $\lambda \in \Lambda$, let $\mu_\lambda \in \Delta(\Lambda)$ be a probability distribution over labels describing how label-λ agents are matched. The last component of a matching profile is then the vector $\mu = (\mu_\lambda)_{\lambda\in\Lambda}$, which satisfies the consistency condition:

$$p[\lambda]\mu_\lambda[\lambda'] = p[\lambda']\mu_{\lambda'}[\lambda] \quad \text{for all } \lambda, \lambda' \in \Lambda.$$

Given a matching profile (Λ, p, q, μ), for each $\lambda, \lambda' \in \Lambda$, we let $\gamma_{\lambda,\lambda'} \in \Delta(\mathcal{A})$ describe the conditional distribution of agreements reached across matches between label-λ and label-λ' agents. An **agreement profile** (with incomplete information) $\gamma = (\gamma_{\lambda,\lambda'})$ is a collection of such conditional distributions satisfying the same exchangeability condition as in the case of complete information. We assume that agreements in the current outcome are observable and that any information they reveal is already encoded in the labels. Formally, conditional on the matched agents' labels, the observed agreement conveys no additional information about their preference types. Thus, for any $\lambda, \lambda' \in \Lambda$, observing any agreement in the support of $\gamma_{\lambda,\lambda'}$ leaves the posterior type distribution of the label-λ agent equal to q_λ and, analogously, that of the label-λ' agent equal to $q_{\lambda'}$.[23]

As before, the combination of a matching profile and an agreement profile $(\Lambda, p, q, \mu, \gamma)$ is called an **outcome** (with incomplete information). As is standard in the literature on matching under incomplete information, we assume that all agents in the population correctly understand the status quo outcome.

Remark 3. Because the informational component (Λ, p, q) is itself part of the outcome, information is *endogenous* under the notion of stability defined below (see also, e.g., Liu et al., 2014; Liu, 2020; Chen and Hu, 2020, 2023; Liu, 2025; Wang, 2026). In this paper, we impose no exogenous informational restrictions and consider the set of all outcomes with labels that encode arbitrary information. Such restrictions can nevertheless be incorporated by specifying, for example, a commonly understood signal structure and restricting attention to outcomes consistent with it.[24] All

[23] If observing an agreement induced a posterior over a label-λ agent's type different from q_λ, the label system could be refined so that the resulting posterior corresponded to a separate label. The only restriction imposed by this representation is therefore the finiteness of Λ, which limits the granularity of the information encoded by labels. A fully general formulation could allow a continuum of labels, but doing so would add technical complexity without offering additional economic insights.

[24] For illustration, consider a population state $(\theta, \tau, \varepsilon)$ and suppose that information about agents' types is generated

of our results continue to hold under such exogenous informational restrictions, provided they do not fully disclose all information—that is, as long as some incomplete information remains.

4.1 Stable Outcome with Incomplete Information

For $t \in \{\theta, \tau\}$ and $\lambda \in \Lambda$, we write

$$\mathbb{E}_{q_\lambda}[u_t(a, a', \cdot)] = u_t(a, a', \theta) q_\lambda[\theta] + u_t(a, a', \tau) q_\lambda[\tau],$$

which is the expected utility of a type-t agent from playing (a, a') with a label-λ partner. For $\lambda, \lambda' \in \Lambda$, let $CE_{\lambda,\lambda'}$ denote the set of incomplete-information correlated equilibria between label-λ and label-λ' agents; formally, $\alpha \in CE_{\lambda,\lambda'}$ if and only if

$$f \in \arg\max_g \mathbb{E}_\alpha[\mathbb{E}_{q_{\lambda'}}[u_t(g(a), a', \cdot)]] \text{ for each } t \in \text{supp}(q_\lambda)$$
$$f \in \arg\max_g \mathbb{E}_{\rho(\alpha)}[\mathbb{E}_{q_\lambda}[u_{t'}(g(a), a', \cdot)]] \text{ for each } t' \in \text{supp}(q_{\lambda'}).$$

In words, when facing a label-λ' partner, each label-λ agent finds it optimal to follow the agreement, given the belief $q_{\lambda'}$ about the partner's underlying type. The second condition imposes the analogous obedience requirement on every type of the label-λ' agent.[25]

Definition 7. With incomplete information, an agreement profile γ is a **correlated equilibrium profile** with respect to μ if for every $\lambda, \lambda' \in \Lambda$ such that $\mu_\lambda[\lambda'] > 0$, $\gamma_{\lambda,\lambda'}[CE_{\lambda,\lambda'}] = 1$.

Because agents observe only the labels of potential partners, rather than their preference types, we require an additional definition to properly formulate pairwise deviations under incomplete information.

Definition 8. For a label $\lambda \in \Lambda$, a tuple (λ, D, d) is a **deviation plan** if (i) D is a nonempty subset of $\text{supp}(q_\lambda)$ and (ii) $d : D \to A^A$.

In words, D is the set of preference types among label-λ agents that participate in a pairwise deviation, while d specifies their behavior. For each $t \in D$, $d(t) : A \to A$ is the strategy used in

before matching according to a signal structure (ξ_θ, ξ_τ). For an agent of type $t \in \{\theta, \tau\}$, a publicly observed signal that perfectly reveals her preference type is generated with probability $\xi_t > 0$; with the remaining probability $1 - \xi_t$, no signal is observed. An outcome $(\Lambda, p, q, \mu, \gamma)$ consistent with this signal structure must satisfy the following additional conditions: there exist $\lambda, \lambda' \in \Lambda$ such that $q_\lambda[\theta] = 1$, $q_{\lambda'}[\tau] = 1$, $p[\lambda] \geq (1 - \varepsilon)\xi_\theta$, and $p[\lambda'] \geq \varepsilon\xi_\tau$. Intuitively, these conditions require the outcome to preserve the information disclosed by the initial signal structure while allowing labels to endogenously reveal additional information.

[25] This notion differs slightly from the standard notion of correlated equilibrium under incomplete information (see, e.g., Forges, 1993, 2006; Liu, 2015; Bergemann and Morris, 2016). In our setting, agents observe and understand the agreement played in their match, and any information conveyed by that observation is already incorporated into their beliefs through the labels. In this respect, our notion of incomplete-information correlated equilibrium resembles the rational expectations equilibrium studied by Koh (2023).

the deviation: upon receiving recommendation a, a type-t agent takes action $d(t)(a)$. As in the complete-information case, we call a pair $(\bar{\lambda}, \bar{\alpha}) \in \Lambda \times \mathcal{A}$ a **current position** of a label-λ agent under the outcome $(\Lambda, p, q, \mu, \gamma)$ if $\mu_\lambda[\bar{\lambda}] > 0$ and $\bar{\alpha} \in \text{supp}(\gamma_{\lambda,\bar{\lambda}})$.

Definition 9. Fix an outcome with incomplete information $(\Lambda, p, q, \mu, \gamma)$. An **incomplete-information blocking pair** exists if there are types $t, t' \in \{\theta, \tau\}$; labels $\lambda, \lambda' \in \Lambda$ with $q_\lambda[t] > 0$ and $q_{\lambda'}[t'] > 0$; current positions $(\bar{\lambda}, \bar{\alpha})$ of label λ and $(\bar{\lambda}', \bar{\alpha}')$ of label λ'; and an agreement $\hat{\alpha} \in \mathcal{A}$ such that for any deviation plans (λ, D, d) and (λ', D', d') satisfying $t \in D$, $t' \in D'$, $d(t) = f$, and $d'(t') = f$, we have

(i) $f \in \arg\max_g \mathbb{E}_{\hat{\alpha}}[\mathbb{E}_{q_{\lambda'}}[u_t(g(a), d'(\cdot)(a'), \cdot) \mid D']]$ and
$f \in \arg\max_g \mathbb{E}_{\rho(\hat{\alpha})}[\mathbb{E}_{q_\lambda}[u_{t'}(g(a), d(\cdot)(a'), \cdot) \mid D]]$;

(ii) $\mathbb{E}_{\hat{\alpha}}[\mathbb{E}_{q_{\lambda'}}[u_t(a, d'(\cdot)(a'), \cdot) \mid D']] > \mathbb{E}_{\bar{\alpha}}[\mathbb{E}_{q_{\bar{\lambda}}}[u_t(a, a', \cdot)]]$ and
$\mathbb{E}_{\rho(\hat{\alpha})}[\mathbb{E}_{q_\lambda}[u_{t'}(a, d(\cdot)(a'), \cdot) \mid D]] > \mathbb{E}_{\bar{\alpha}'}[\mathbb{E}_{q_{\bar{\lambda}'}}[u_{t'}(a, a', \cdot)]]$.

In the definition above, for an agent of type $t \in \{\theta, \tau\}$, a pair of action recommendations (a, a'), and a deviation plan (λ', D', d') of the deviating partner, the conditional expected utility $\mathbb{E}_{q_{\lambda'}}[u_t(g(a), d'(\cdot)(a'), \cdot) \mid D']$ is evaluated using the probability distribution $q_{\lambda'}$ conditional on the partner's type belonging to D'. If D' is a singleton, then the expectation is degenerate.

Definition 9 describes a situation in which a pair of agents, despite observing only each other's labels, can still reach an agreement and carry out a mutually beneficial deviation.[26] In particular, a blocking pair consists of a type-t agent with label λ (i.e., a type-t_λ agent) and a type-t' agent with label λ' (i.e., a type-$t'_{\lambda'}$ agent). We call them the **targeted** agents. Provided that the targeted type-$t'_{\lambda'}$ agent participates and follows $\hat{\alpha}$, the type-t_λ agent also finds it optimal to follow $\hat{\alpha}$ and thereby obtains a strict utility improvement, regardless of whether or how nontargeted agents with label λ' participate in the deviation. The same argument applies symmetrically to the targeted type-$t'_{\lambda'}$ agent. Thus, each targeted agent's incentive to deviate is guaranteed conditional on the participation and obedience of her targeted partner.[27]

Remark 4. In Definition 9, a deviating agent relies on the rationality of her targeted partner, conditional on her own participation. While it is possible to further account for the rationality of nontargeted types, we adopt a more conservative approach for two reasons. First, rational behavior of nontargeted agents may reveal additional information in a deviation and may induce the deviating

[26] Our approach is akin to the cheap talk paradigm, in which information is conveyed through endogenous behavior. This is fundamentally different from models in which agents incur costs to signal their types or increase the likelihood of being identified.

[27] Consider the following statement a type-t_λ agent could make to an agent with label λ': "I am of type t, and I promise to participate in the deviation and follow the recommendation of $\hat{\alpha}$ for me. I ask you to join the deviation and follow the recommendation of $\hat{\alpha}$ for you if you are of type t'. Why should you follow my advice? Because as long as I fulfill my promise, following the recommendation of $\hat{\alpha}$ is optimal for you, which strictly improves your expected utility. Why should you trust my promise? Because as long as you follow my advice, following the recommendation of $\hat{\alpha}$ is optimal for me, which strictly improves my expected utility."

partner to adjust her behavior after the rematch.[28] Accounting for this possibility would require us to take a stance on how agents anticipate and respond to the resulting inferences and behavioral adjustments; see a related discussion in Liu (2020). Our definition avoids this issue because the responses of nontargeted agents—whether rational or irrational—do not affect the decisions of the targeted agents. Second, a more demanding criterion for blocking yields a weaker notion of stability, thereby strengthening the positive result in Proposition 6. At the same time, the negative result in Proposition 5 does not rely on irrational behavior by any nontargeted type; see footnotes 32 and 33.

We use an example to illustrate the notion of an incomplete-information blocking pair.

Example 5. Consider the prisoners' dilemma material game as follows:

	x	y
x	$4,4$	$1,5$
y	$5,1$	$3,3$

Suppose that type-θ agents have efficient preferences $u_\theta(a,a',t) = \pi(a,a') + \pi(a',a)$, while type-$\tau$ agents are selfish $u_\tau(a,a',t) = \pi(a,a')$. Consider a population state $(\theta,\tau,\varepsilon)$ and an outcome (Λ,p,q,μ,γ) as follows. Each type constitutes one half of the population, so $\varepsilon = 1/2$. The matching profile (Λ,p,q,μ) satisfies $\Lambda = \{\lambda\}$, $p[\lambda] = 1$, $q_\lambda[\theta] = q_\lambda[\tau] = 1/2$, and $\mu_\lambda[\lambda] = 1$. The agreement profile γ satisfies $\gamma_{\lambda,\lambda}[\alpha] = 1$, where $\alpha[y,y] = 1$. This outcome is depicted in Figure 1 below.

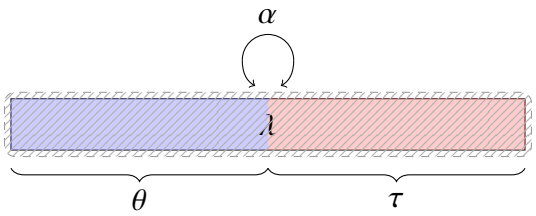

Figure 1: The outcome in Example 5.

To show that an incomplete-information blocking pair exists, consider two type-θ_λ agents who target one another and propose the efficient agreement $\hat{\alpha}$, where $\hat{\alpha}[x,x] = 1$. By symmetry, it suffices to verify the incentives of one of these agents. Consider any deviation plan (λ,D,d) of her deviating partner satisfying $\theta \in D$ and $d(\theta) = f$; that is, the targeted type-θ_λ partner participates in the deviation and follows $\hat{\alpha}$. We show that following $\hat{\alpha}$ is optimal for the agent and makes her strictly better off. There are two cases to consider:

- Suppose $\tau \notin D$. Since x is a best response against x for a type-θ agent, following $\hat{\alpha}$ is optimal. Moreover, following $\hat{\alpha}$ yields a utility of 8 to the deviating type-θ agent, exceeding her status quo expected utility of 6.

[28] This occurs, for example, when the deviation plan (λ,D,d) satisfies $D = \{\theta,\tau\}$ and $d(\theta) \neq d(\tau)$.

- Suppose $\tau \in D$. Because $\hat{\alpha}$ recommends x with probability one, only the action $d(\tau)(x)$ of the deviating partner matters. If $d(\tau)(x) = x$, the partner plays x regardless of her type. Following $\hat{\alpha}$ is therefore optimal and gives the type-θ agent expected utility $8 > 6$. If $d(\tau)(x) = y$, following $\hat{\alpha}$ is still optimal because $8(1/2) + 6(1/2) = 7 > 6 = 6(1/2) + 6(1/2)$, which also makes her strictly better off.

Therefore, conditional on a targeted type-θ_λ partner participating in the deviation and following $\hat{\alpha}$, a type-θ_λ agent also finds it optimal to follow $\hat{\alpha}$ and obtains a strict utility improvement. This conclusion remains valid even if nontargeted type-τ_λ agents join the deviation and behave arbitrarily. Hence, an incomplete-information blocking pair exists. ◊

We extend the notion of stable outcome to the case of incomplete information.

Definition 10. An outcome with incomplete information $(\Lambda, p, q, \mu, \gamma)$ is **incomplete-information stable** if it satisfies:

(i) (Internal stability) γ is a correlated equilibrium profile with respect to μ;

(ii) (External stability) There is no incomplete-information blocking pair.

When $\Lambda = \{\lambda, \lambda'\}$, $p[\lambda] = 1 - \varepsilon$, $p[\lambda'] = \varepsilon$, and $q_\lambda[\theta] = q_{\lambda'}[\tau] = 1$, agents' types are fully revealed, and Definitions 7 and 9 then coincide with Definitions 1 and 2, respectively. Hence, incomplete-information stability reduces to complete-information stability. The existence of an incomplete-information stable outcome is therefore guaranteed. An outcome that is stable under incomplete information, however, need not remain stable if agents' preference types become fully observable, as full disclosure may create additional blocking opportunities. As highlighted in Remark 3, information in a stable outcome is *endogenous*: the degree to which preference types are revealed depends on agents' matching patterns and behavior.

4.2 Evolutionary Dominance with Incomplete Information

Given an incomplete-information stable outcome $(\Lambda, p, q, \mu, \gamma)$ in population state $(\theta, \tau, \varepsilon)$, the average material payoffs for agents of type θ and type τ are given by

$$G_\theta(\Lambda, p, q, \mu, \gamma) = \frac{1}{1-\varepsilon} \sum_{\lambda \in \Lambda} \left\{ \sum_{\lambda' \in \Lambda} \left[\int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a, a')] \, d\gamma_{\lambda,\lambda'}[\alpha] \right] \mu_\lambda[\lambda'] \right\} p[\lambda] q_\lambda[\theta],$$

$$G_\tau(\Lambda, p, q, \mu, \gamma) = \frac{1}{\varepsilon} \sum_{\lambda \in \Lambda} \left\{ \sum_{\lambda' \in \Lambda} \left[\int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a, a')] \, d\gamma_{\lambda,\lambda'}[\alpha] \right] \mu_\lambda[\lambda'] \right\} p[\lambda] q_\lambda[\tau].$$

Our notions of evolutionary dominance can be naturally extended to incorporate incomplete information.

Definition 11. Fix a material game $\mathcal{G}$. With incomplete information,

(i) A preference type $\theta \in \Theta$ is **evolutionarily dominant** if for every $\tau \in \Theta$ and every $\varepsilon \in (0,1)$, every incomplete-information stable outcome $(\Lambda, p, q, \mu, \gamma)$ in the population state $(\theta, \tau, \varepsilon)$ satisfies $G_\theta(\Lambda, p, q, \mu, \gamma) \geq G_\tau(\Lambda, p, q, \mu, \gamma)$.

(ii) A preference type $\theta \in \Theta$ is **evolutionarily dominated** if there exists another type $\tau \in \Theta$ such that for every $\varepsilon \in (0,1)$, every incomplete-information stable outcome $(\Lambda, p, q, \mu, \gamma)$ in the population state $(\theta, \tau, \varepsilon)$ satisfies $G_\theta(\Lambda, p, q, \mu, \gamma) \leq G_\tau(\Lambda, p, q, \mu, \gamma)$, with strict inequality for some incomplete-information stable outcome.

A morally constrained selfish type is evolutionarily dominant under complete information (Proposition 1) because agents of such a type can secure a material payoff of $m/2$ by rematching with one another to play a fair and efficient agreement. This mechanism, however, relies critically on complete information. When morally constrained selfish agents can no longer identify others of their own type, they may be reluctant to undertake such deviations, causing the type to lose its evolutionary advantage. The following example illustrates this possibility.

Example 6. Consider again the material game from Example 2, reproduced below:

	x	y
x	$1, 1$	$4, 2$
y	$2, 4$	$1, 1$

Let θ denote a morally constrained selfish type. Let τ denote a type with efficient preferences when matched with its own type but for which x is strictly dominant otherwise:

$$u_\tau(a, a', t) = \begin{cases} \pi(a, a') + \pi(a', a) & \text{if } t = \tau, \\ 8 \cdot \mathbf{1}_{\{a=x\}} & \text{if } t \neq \tau. \end{cases}$$

Now consider a population state $(\theta, \tau, \varepsilon)$ and an outcome $(\Lambda, p, q, \mu, \gamma)$ as follows. Assume that the proportion of type-τ agents satisfies $\varepsilon > 3/4$. The matching profile (Λ, p, q, μ) satisfies $\Lambda = \{\lambda, \lambda'\}$, $p[\lambda] = p[\lambda'] = 1/2$, $q_\lambda[\theta] = 2(1-\varepsilon) < 1/2$, $q_{\lambda'}[\tau] = 1$, and $\mu_\lambda[\lambda'] = \mu_{\lambda'}[\lambda] = 1$. The agreement profile γ satisfies $\gamma_{\lambda,\lambda'}[\alpha] = 1$, where $\alpha[y, x] = 1$. This outcome is depicted in Figure 2.

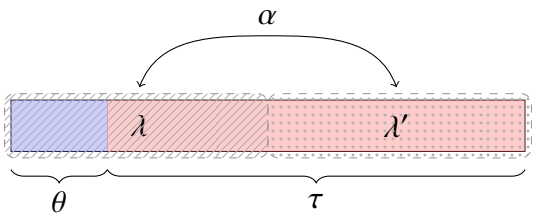

Figure 2: The outcome in Example 6.

We formally verify in Appendix A.2.1 that $(\Lambda, p, q, \mu, \gamma)$ is incomplete-information stable. For intuition, suppose type-θ_λ agents target one another and propose, for example, a fair and efficient agreement $\hat{\alpha}$ satisfying $\hat{\alpha}[x, y] = \hat{\alpha}[y, x] = 1/2$. Although both sides would follow $\hat{\alpha}$ and gain under complete information, nontargeted type-τ_λ agents may join the deviation and play x under incomplete information, leaving the targeted agents with expected utility $((2+4)/2)q_\lambda[\theta] + ((2+0)/2)(1 - q_\lambda[\theta]) = 1 + 2q_\lambda[\theta] < 2$, below their status quo utility. Thus, the deviation is not viable under incomplete information. Notably, although α is efficient, type-θ agents earn a strictly lower average material payoff than type-τ agents in this incomplete-information stable outcome. ◊

Example 6 constructs a particular incomplete-information stable outcome in which the morally constrained selfish type earns a lower average material payoff than the competing type. The following proposition shows that the underlying vulnerability is more general: whenever efficient play is unique up to role reversal and yields unequal material payoffs, a morally constrained selfish type is evolutionarily dominated under incomplete information.[29]

Proposition 5. *With incomplete information, if* $\mathcal{E} = \{(a^*, a^\dagger), (a^\dagger, a^*)\}$ *and* $\pi(a^*, a^\dagger) \neq \pi(a^\dagger, a^*)$, *then a morally constrained selfish type is evolutionarily dominated.*

The proof constructs a competing type with two complementary properties. First, regardless of its population share, this type earns an average material payoff of at least $m/2$ in every incomplete-information stable outcome. This ensures that it always weakly outperforms the morally constrained selfish type (Lemma 4). Second, its behavior sometimes deters morally constrained selfish agents from escaping the disadvantaged side of efficient play, allowing the payoff inequality to be strict in some stable outcome (Lemma 5). Together, these properties establish evolutionary domination.

In contrast, the next proposition shows that a parochial efficient type stands out even with incomplete information.

Proposition 6. *With incomplete information, for any material game* $\mathcal{G}$, *a parochial efficient type is evolutionarily dominant.*

To understand the sharp contrast between Propositions 5 and 6, it is helpful to consider the underlying logic of Example 6. There, morally constrained selfish agents cannot escape an unfavorable position through pairwise deviations because nontargeted type-τ_λ agents may join and behave in ways that make the proposed rematch unprofitable. Parochial efficient agents, by contrast, place first priority on matching with their own type, so the behavior of nontargeted type-τ_λ agents cannot eliminate the gains from rematching. They can therefore break away from the same disadvantageous position.

[29] This class encompasses many familiar games in which efficient play requires agents to assume different roles, with one role yielding a larger material payoff than the other. Canonical examples include the Battle of the Sexes (Luce and Raiffa, 1957) (with a suitably redefined action set), the Leader game (Rapoport, 1967), and the two-player volunteer's dilemma (Diekmann, 1985). Even outside this class, simple examples can be constructed in which a morally constrained selfish type fails to be evolutionarily dominant.

Example 7 (Example 6 revisited)**.** Consider the population state and outcome from Example 6, except that θ now denotes a parochial efficient type. Type-θ_λ agents derive utility 0 in the status quo. Thus, two type-θ_λ agents can target one another and propose the efficient agreement $\hat{\alpha}$ from Example 6. If the nontargeted type-τ_λ agents do not participate, each targeted agent finds it optimal to follow $\hat{\alpha}$ and obtains expected utility $6 > 0$. If they also participate, following $\hat{\alpha}$ remains optimal for each targeted agent and yields expected utility $6q_\lambda[\theta] + 0(1 - q_\lambda[\theta]) > 0$, regardless of how the nontargeted agents behave. Thus, the proposed deviation constitutes an incomplete-information blocking pair. ◊

The proof of Proposition 6 shows how parochial efficient agents overcome informational frictions. Specifically, when a parochial efficient type is present, incomplete-information stable outcomes satisfy two key properties (see Lemmas 2 and 3). First, matching is perfectly assortative by label: matched partners always carry the same label. This, on average, limits the competing type's ability to gain from any material-payoff sacrifice made by the parochial efficient type. Second, for every label λ such that $q_\lambda[\theta] > 0$, agents carrying that label interact efficiently with one another. Together, these properties ensure that a parochial efficient type attains an average material payoff at least as high as that of any competing type.

5 Polymorphism

Thus far, we have focused on evolutionary dominance of a single type. In this section, we extend the framework to polymorphic populations, in which several preference types coexist. For simplicity, we maintain complete information in this section; the corresponding extension under incomplete information is conceptually similar.

Let $\nu \in \Delta(\Theta)$ be a **population distribution** with finite support $\Theta_\nu = \mathrm{supp}(\nu)$. For each type $t \in \Theta_\nu$, the mass of type-t agents in the population is denoted by $\nu[t] > 0$. An outcome (μ, γ) is defined as in the two-type case. The matching profile is now a vector $\mu = (\mu_t)_{t\in\Theta_\nu}$, where $\mu_t \in \Delta(\Theta_\nu)$ gives the distribution of partner types for type-t agents. In particular, $\mu_t[t']$ is the fraction of type-t agents matched with type-t' agents. The matching profile satisfies the consistency condition

$$\nu[t]\mu_t[t'] = \nu[t']\mu_{t'}[t] \qquad \text{for all } t, t' \in \Theta_\nu.$$

For each $t, t' \in \Theta_\nu$, let $\gamma_{t,t'} \in \Delta(\mathcal{A})$ denote the conditional distribution of agreements reached in matches between type-t and type-t' agents. As before, the agreement profile $\gamma = (\gamma_{t,t'})$ satisfies the exchangeability condition.

The definitions of a complete-information blocking pair and a complete-information stable outcome under a population distribution ν extend directly by replacing $\{\theta, \tau\}$ with Θ_ν in Definitions 2 and 3. The resulting environment is a one-sided matching market with contracts and a continuum of agents. An important recent literature studies existence and structural properties of stable outcomes

in related large matching markets (Che et al., 2019; Greinecker and Kah, 2021; Jagadeesan and Vocke, 2024; Carmona and Laohakunakorn, 2024). Adapting arguments from this literature, we establish the following existence result.

Proposition 7. *For any material game $\mathcal{G}$ and any population distribution ν, a complete-information stable outcome exists.*

Given a complete-information stable outcome (μ, γ) under a population distribution ν, for each $t \in \Theta_\nu$, the average material payoff earned by type-t agents is given by

$$G_t(\mu, \gamma) = \sum_{t' \in \Theta_\nu} \left[\int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a, a')] \, d\gamma_{t,t'}[\alpha] \right] \mu_t[t'].$$

We next follow Dekel et al. (2007) and ask whether a population distribution can resist invasion by any competing preference type. We assume that mutations are sufficiently rare and happen one at a time, with the population able to fully adjust before the next mutation occurs (Weibull, 1995). Accordingly, we treat the population distribution as incumbent and restrict the mutant to a sufficiently small population share. This leads to the following local notion of evolutionary dominance.

Definition 12. A population distribution ν is **locally evolutionarily dominant** if there exists $\bar{\varepsilon} > 0$ such that for every mutant type $t' \in \Theta$, $\varepsilon \in (0, \bar{\varepsilon})$, and complete-information stable outcome $(\tilde{\mu}, \tilde{\gamma})$ under the post-entry population distribution $\tilde{\nu} = (1 - \varepsilon)\nu + \varepsilon\delta_{t'}$, we have $G_t(\tilde{\mu}, \tilde{\gamma}) \geq G_{t'}(\tilde{\mu}, \tilde{\gamma})$ for every $t \in \Theta_\nu$.[30]

Thus, whenever the mutant's population share lies below a common threshold, every incumbent type must earn at least as much as the mutant in every post-entry stable outcome. No sufficiently small invasion can therefore displace any type in the incumbent population. Under this definition, the mutant type may already belong to the incumbent support, in which case the perturbation changes only its population share.

We now derive necessary conditions that describe crucial properties of population distributions that are locally evolutionarily dominant.

Proposition 8. *Suppose that the population distribution ν is locally evolutionarily dominant. Then every complete-information stable outcome (μ, γ) under ν satisfies the following:*

(i) $G_t(\mu, \gamma) = G_{t'}(\mu, \gamma)$ *for all* $t, t' \in \Theta_\nu$.

(ii) For any $t, t' \in \Theta_\nu$ *such that* $\mu_t[t'] > 0$, *every agreement* $\alpha \in \operatorname{supp}(\gamma_{t,t'})$ *is efficient. Moreover, either* $t = t'$ *or* $\mathbb{E}_\alpha[\pi(a, a')] = \mathbb{E}_{\rho(\alpha)}[\pi(a, a')]$.

Part (i) of Proposition 8 means that a locally evolutionarily dominant ν should itself be *balanced*: all types in its support must earn the same average material payoff. Otherwise, natural selection

[30] We write $\delta_{t'} \in \Delta(\Theta)$ for the Dirac measure that assigns probability one to type t'.

would alter their relative frequencies even in the absence of mutations. Whereas Dekel et al. (2007) impose balancedness as part of their definition of evolutionary stability for a polymorphic population, here it follows from local evolutionary dominance.

Part (ii) requires efficient play in every complete-information stable outcome. If some incumbents play inefficiently, a parochial efficient mutant can separate from the incumbent population and play efficiently with its own type, resulting in a post-entry stable outcome in which the mutant outperforms at least one incumbent type. Part (ii) also imposes a form of payoff symmetry on cross-type matches: whenever two distinct types are matched, the two sides must receive the same expected material payoff.

While the necessary conditions in Proposition 8 may appear demanding, the following result provides a sufficient condition.

Proposition 9. *If Θ_ν consists of morally constrained selfish and parochial efficient types, then ν is locally evolutionarily dominant.*

The result follows from the logic of Propositions 1 and 2, and we therefore omit a formal proof. In any stable outcome following the entry of an arbitrary mutant, all morally constrained selfish incumbents earn at least $m/2$, while parochial efficient incumbents match with their own type and play efficiently, earning $m/2$ on average. Because the aggregate material payoff of the population is bounded above by $m/2$, the mutant earns at most this bound on average. Thus, all incumbent types weakly outperform the mutant.

It is natural to ask whether a locally evolutionarily dominant population can contain preference types beyond the two kinds considered above. Of particular interest are mutually *heterophilic* types—distinct types that prefer to match with one another. Such preferences may arise when social identities or observable characteristics shape partner choice and behavior. The following example shows that heterophily can indeed be sustained.

Example 8. Consider again the prisoners' dilemma material game in Example 5, which is reproduced below:

	x	y
x	$4, 4$	$1, 5$
y	$5, 1$	$3, 3$

Let ν denote a population distribution that contains two types $\Theta_\nu = \{\theta, \tau\}$ and $\nu[\theta] = \nu[\tau] = 1/2$. The utility functions of θ and τ are given by, respectively,

$$u_\theta(a, a', t) = \pi(a, a') + \pi(a', a) \cdot \mathbf{1}_{\{t=\tau\}} \quad \text{and} \quad u_\tau(a, a', t) = \pi(a, a') + \pi(a', a) \cdot \mathbf{1}_{\{t=\theta\}}$$

We now show that ν is locally evolutionarily dominant. Consider a mutant type t', an $\varepsilon > 0$ sufficiently small, and an arbitrary complete-information stable outcome $(\tilde{\mu}, \tilde{\gamma})$ under the post-entry population $\tilde{\nu} = (1 - \varepsilon)\nu + \varepsilon\delta_{t'}$. There are two cases:

- If $t' \notin \Theta_\nu$, then we must have $\tilde{\mu}_\theta[\tau] = \tilde{\mu}_\tau[\theta] = 1$ and $\tilde{\gamma}_{\theta,\tau}[\hat{\alpha}] = 1$, where $\hat{\alpha}[x, x] = 1$. It follows that $G_\theta(\tilde{\mu}, \tilde{\gamma}) = G_\tau(\tilde{\mu}, \tilde{\gamma}) = 4 \geq G_{t'}(\tilde{\mu}, \tilde{\gamma})$.
- If $t' \in \Theta_\nu$, suppose $t' = \theta$ without loss. Then all type-τ agents match with type-θ agents and play (x, x), while excess type-θ agents must match among themselves. Therefore, $\tilde{\mu}_\theta[\theta] = 2\varepsilon/(1+\varepsilon)$, $\tilde{\mu}_\theta[\tau] = (1-\varepsilon)/(1+\varepsilon)$, and $\tilde{\mu}_\tau[\theta] = 1$; moreover, $\tilde{\gamma}_{\theta,\tau}[\hat{\alpha}] = 1$ and $\tilde{\gamma}_{\theta,\theta}[\tilde{\alpha}] = 1$, where $\tilde{\alpha}[y, y] = 1$. Consequently, $G_\theta(\tilde{\mu}, \tilde{\gamma}) < 4 = G_\tau(\tilde{\mu}, \tilde{\gamma})$. Both incumbent types weakly outperform the entrant.

Therefore, the population distribution ν that consists of two heterophilic types is locally evolutionarily dominant. Interestingly, although both types would play inefficiently with same-type partners, this does not destabilize the population because such matches do not occur at the balanced composition. ◇

6 Concluding Remarks

This paper studies preference evolution under endogenous matching by combining stable partner choice with equilibrium play. Evolutionary success requires agents both to sustain efficient play and to protect themselves from exploitation. We identify two classes of preference types that combine these forces: morally constrained selfish types and parochial efficient types. Both classes are evolutionarily dominant under complete information. Under incomplete information, however, morally constrained selfish types can be dominated, whereas parochial efficient types remain dominant. By prioritizing same-type matching, parochialism acts as a form of group-level selfishness and enables agents to overcome the informational frictions that impede assortative matching.

Although Charness and Rabin (2002) and Engelmann and Strobel (2004) find behavior consistent with efficiency concerns, experimental support for an unconditional preference for efficiency remains inconclusive. Our analysis suggests that efficient behavior may reflect more structured motives: agents may either maximize their own material payoffs subject to an efficiency constraint or value efficiency only when interacting with their own type. Standard experimental designs based on exogenous matching are therefore not well suited to detect and distinguish these motives. A more informative design would allow subjects to communicate their intended play and form pairs by mutual consent.[31] Once matching patterns and play have stabilized, researchers could, for example, elicit subjects' other-regarding preferences separately toward current partners and nonpartners. Greater weight on efficiency in choices involving current partners than in otherwise identical choices involving nonpartners would provide evidence consistent with parochial efficiency.

[31] An experimental literature beginning with Ehrhart and Keser (1999) and Hauk and Nagel (2001) studies behavior under endogenous partner choice. A common finding is that partner choice promotes cooperation and coordination on efficient outcomes by allowing like-minded subjects to associate. Related experiments find similar effects when subjects are grouped assortatively by elicited cooperativeness (Coricelli et al., 2004; Burlando and Guala, 2005; de Oliveira et al., 2015).

Several important questions remain for future research. First, a dynamic model could replace our reduced-form notion of stability with history-dependent matching and play. The analysis of repeated coalitional behavior by Ali and Liu (2026) suggests that agents with perfectly aligned preferences attain a jointly optimal outcome in every dynamically stable process. Because the interests of two parochial efficient agents are indeed perfectly aligned, evolutionary dominance of such types may extend to a genuinely dynamic setting. Second, preferences and institutions may coevolve. Institutions shape incentives for matching and play through taxes and subsidies (Hiller and Touré, 2021), property-rights protection (Bisin and Verdier, 2021), religious infrastructure (Bisin et al., 2021), and integration policies (Wu, 2017). Extending our framework along the lines of Bisin and Verdier (2024) would allow a joint analysis of the evolution of preferences and institutions.

References

Akdeniz, A. and M. van Veelen (2021): "The evolution of morality and the role of commitment," *Evolutionary Human Sciences*, 3, e41.

——— (2023): "Evolution and the ultimatum game," *Games and Economic Behavior*, 142, 570–612.

Alger, I. (2023): "Evolutionarily Stable Preferences," *Philosophical Transactions of the Royal Society B*, 378, 20210505.

Alger, I. and L. Lehmann (2023): "Evolution of semi-Kantian preferences in two-player assortative interactions with complete and incomplete information and plasticity," *Dynamic games and Applications*, 13, 1288–1319.

Alger, I. and J. W. Weibull (2013): "Homo Moralis: Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 81, 2269–2302.

——— (2016): "Evolution and Kantian morality," *Games and Economic Behavior*, 98, 56–67.

——— (2019): "Evolutionary Models of Preference Formation," *Annual Review of Economics*, 11, 329–354.

Alger, I., J. W. Weibull, and L. Lehmann (2020): "Evolution of preferences in structured populations: Genes, guns, and culture?" *Journal of Economic Theory*, 185, 104951.

Ali, S. N. and C. Liu (2026): "Coalitions in Repeated Games," *Econometrica*, forthcoming.

Aumann, R. J. (1974): "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, 1, 67–96.

——— (1987): "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1–18.

Avataneo, M. (2026): "The Evolutionary Success of Moral Universalism vs Moral Particularism," Working paper.

Avataneo, M., T. Norman, and N. Persico (2025): "The Evolutionary Stability of Moral Foundations," *The Quarterly Journal of Economics*, 140, 2459–2506.

Bandhu, S. and R. Lahkar (2023): "Survival of altruistic preferences in a large population public goods game," *Economics Letters*, 226, 111113.

Becker, G. S. (1976): "Altruism, Egoism, and Genetic Fitness," *Journal of Economic Literature*, 14, 817–826.

Bergemann, D. and S. Morris (2016): "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," *Theoretical Economics*, 11, 487–522.

Bernhard, H., U. Fischbacher, and E. Fehr (2006): "Parochial altruism in humans," *Nature*, 442, 912–915.

Bisin, A., J. Rubin, A. Seror, and T. Verdier (2021): "Culture, Institutions & the Long Divergence," *NBER Working Papers 28488 National Bureau of Economic Research, Inc.*

Bisin, A. and T. Verdier (2021): "Phase Diagrams in Historical Economics: Culture and Institutions," in *Handbook of Historical Economics*, ed. by A. Bisin and G. Federico, Amsterdam: Elsevier North Holland.

——— (2024): "On the Joint Evolution of Culture and Political Institutions: Elites and Civil Society," *Journal of Political Economy*, 132, 1485–1564.

Burlando, R. M. and F. Guala (2005): "Heterogeneous Agents in Public Goods Experiments," *Experimental Economics*, 8, 35–54.

Carmona, G. and K. Laohakunakorn (2024): "Stable Matching in Large Markets with Occupational Choice," *Theoretical Economics*, 19, 1261–1304.

Catonini, E. and Z. Wang (2026): "Interim Agreements in Matching with Incomplete Information," Working paper.

Charness, G. and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117, 817–869.

Che, Y.-K., J. Kim, and F. Kojima (2019): "Stable Matching in Large Economies," *Econometrica*, 87, 65–110.

Chen, Y.-C. and G. Hu (2020): "Learning by Matching," *Theoretical Economics*, 15, 29–56.

——— (2023): "A Theory of Stability in Matching with Incomplete Information," *American Economic Journal: Microeconomics*, 15, 288–322.

Choi, J. and S. Bowles (2007): "The Coevolution of Parochial Altruism and War," *Science*, 318, 636–640.

Coricelli, G., D. Fehr, and G. Fellner (2004): "Partner Selection in Public Goods Experiments," *The Journal of Conflict Resolution*, 48, 356–378.

de Oliveira, A. C. M., R. T. A. Croson, and C. Eckel (2015): "One bad apple? Heterogeneity and information in public good provision," *Experimental Economics*, 18, 116–135.

Dekel, E., J. C. Ely, and O. Yilankaya (2007): "Evolution of Preferences," *Review of Economics Studies*, 74, 685–704.

Diekmann, A. (1985): "Volunteer's Dilemma," *Journal of Conflict Resolution*, 29, 605–610.

Echenique, F., S. Lee, M. Shum, and B. M. Yenmez (2013): "The Revealed Preference Theory of Stable and Extremal Stable Matchings," *Econometrica*, 81, 153–171.

Ehrhart, K.-M. and C. Keser (1999): "Mobility and Cooperation: On the Run," *Cirano Working paper*.

Ely, J. C. and O. Yilankaya (2001): "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97, 255–272.

Engelmann, D. and M. Strobel (2004): "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments," *American Economic Review*, 94, 857–869.

Forges, F. (1993): "Five Legitimate Definitions of Correlated Equilibrium in Games with Incomplete Information," *Theory and Decision*, 35, 277–310.

——— (2006): "Correlated Equilibrium in Games with Incomplete Information Revisited," *Theory and Decision*, 61, 329–344.

Frank, R. H. (1987): "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review*, 77, 593–604.

Fujiwara-Greve, T. and M. Okuno-Fujiwara (2009): "Voluntarily Separable Repeated Prisoner's Dilemma," *Review of Economic Studies*, 76, 993–1021.

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

Garrido-Lucero, F. and R. Laraki (2021): "Stable Matching Games," *Working paper*.

Graser, C., T. Fujiwara-Greve, J. García, and M. van Veelen (2024): "Repeated Games with Partner Choice," Working paper.

Greinecker, M. and C. Kah (2021): "Pairwise Stable Matching in Large Economies," *Econometrica*, 89, 2929–2974.

Gul, F. and W. Pesendorfer (2016): "Interdependent Preference Models as a Theory of Intentions," *Journal of Economic Theory*, 165, 179–208.

Güth, W. (1995): "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, 24, 323–344.

Güth, W. and H. Kliemt (1998): "Indirect Evolutionary Approach: Bridging the Gap between Rationality and Adaptation," *Rationality and Society*, 10, 377–399.

Güth, W. and M. Yaari (1992): "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game," in *Explaining Process and Change - Approaches to Evolutionary Economics*, ed. by U. Witt, University of Michigan Press, 23–34.

Hamilton, W. D. (1964a): "The Genetical Evolution of Social Behaviour. I," *Journal of Theoretical Biology*, 7, 1–16.

——— (1964b): "The Genetical Evolution of Social Behaviour. II," *Journal of Theoretical Biology*, 7, 17–52.

Hauk, E. and R. Nagel (2001): "Choice of Partners in Multiple Two-Person Prisoner's Dilemma Games: An Experimental Study," *Journal of Conflict Resolution*, 45, 770–793.

Herold, F. and C. Kuzmics (2009): "Evolutionary Stability of Discrimination under Observability," *Games and Economic Behavior*, 67, 542–551.

Hiller, V. and N. Touré (2021): "Endogenous gender power: The two facets of empowerment," *Journal of Development Economics*, 149, 102596.

Hirshleifer, J. (1977): "Economics from a Biological Viewpoint," *Journal of Law and Economics*, 20, 1–52.

Izquierdo, S. S., L. R. Izquierdo, and M. van Veelen (2021): "Repeated Games with Endogenous Separation," Working paper.

Izquierdo, S. S., L. R. Izquierdo, and F. Vega-Redondo (2010): "The Option to Leave: Conditional Dissociation in the Evolution of Cooperation," *Journal of Theoretical Biology*, 267, 76–84.

——— (2014): "Leave and let leave: A sufficient condition to explain the evolutionary emergence of cooperation," *Journal of Economic Dynamics and Control*, 46, 91–113.

Jackson, M. O. (2014): "Networks in the understanding of economic behaviors," *Journal of Economic Perspectives*, 28, 3–22.

Jackson, M. O. and A. Watts (2002): "On the formation of interaction networks in social coordination games," *Games and Economic Behavior*, 41, 265–291.

——— (2010): "Social Games: Matching and the Play of Finitely Repeated Games," *Games and Economic Behavior*, 70, 170–191.

Jagadeesan, R. and K. Vocke (2024): "Stability in Large Markets," *Review of Economic Studies*, 91, 3532–3568.

Koh, P. S. (2023): "Stable Outcomes and Information in Games: An Empirical Framework," *Journal of Econometrics*, 237, 105499.

Lahkar, R. (2019): "Elimination of Non-Individualistic Preferences in Large Population Aggregative Games," *Journal of Mathematical Economics*, 84, 150–165.

Lehmann, L., I. Alger, and J. W. Weibull (2015): "Does evolution lead to maximizing behavior?" *Evolution*, 69, 1858–1873.

Liu, Q. (2015): "Correlation and Common Priors in Games with Incomplete Information," *Journal of Economic Theory*, 157, 49–75.

——— (2020): "Stability and Bayesian Consistency in Two-Sided Markets," *American Economic Review*, 110, 2625–2666.

——— (2025): "A Cooperative Analysis of Incomplete Information," *Working paper*.

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

Luce, R. D. and H. Raiffa (1957): *Games and Decisions: Introduction and Critical Survey*, New York: John Wiley & Sons.

Mill, J. S. (1863): *Utilitarianism*, London: Parker, Son, and Bourn.

Newton, J. (2017): "The preferences of Homo Moralis are unstable under evolving assortativity," *International Journal of Game Theory*, 46, 583–589.

Nowak, M. A., K. M. Page, and K. Sigmund (2000): "Fairness Versus Reason in the Ultimatum Game," *Science*, 289, 1773–1775.

Ok, E. A. and F. Vega-Redondo (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory*, 97, 231–254.

Rapoport, A. (1967): "Exploiter, Leader, Hero, and Martyr: The Four Archetypes of the 2×2 Game," *Behavioral Science*, 12, 81–84.

Robson, A. and L. Samuelson (2011): "The evolutionary foundations of preferences," in *Handbook of social economics*, ed. by J. Benhabib, A. Bisin, and M. Jackson, Elsevier, vol. 1, 221–310.

Rubin, P. and C. Paul (1979): "An Evolutionary Model of Taste for Risk," *Economic Inquiry*, 17(4), 585–596.

Sethi, R. and E. Somanathan (2001): "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97, 273–297.

Wang, Z. (2026): "Rationalizable Stability in Matching with Incomplete Information," *Working paper*.

Weibull, J. W. (1995): *Evolutionary Game Theory*, Cambridge, MA: The MIT Press.

Wu, J. (2017): "Political institutions and the evolution of character traits," *Games and Economic Behavior*, 106, 260–276.

——— (2020): "Labelling, Homophily and Preference Evolution," *International Journal of Game Theory*, 49, 1–22.

Xing, Y. (2020): "Who Shares Risk with Whom and How? Endogenous Matching Selects Risk Sharing Equilibria," Working paper.

A Omitted Proofs

A.1 Proofs for Section 3: Complete Information

A.1.1 Proof of Proposition 1

We first establish the following lemma, which guarantees the existence of an efficient correlated equilibrium between two morally constrained selfish agents.

Lemma 1. *Let θ denote a morally constrained selfish type. There exists an agreement $\tilde{\alpha} \in CE_{\theta,\theta}$ such that $\tilde{\alpha}[\mathcal{E}] = 1$.*

Proof. We choose the following normalization of a morally constrained selfish type,

$$u_\theta(a, a', \theta) = \begin{cases} \pi(a, a'), & \text{if } (a, a') \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases}$$

Let

$$F = \{a \in A : (a, a') \in \mathcal{E} \text{ for some } a' \in A\}$$

and consider the finite symmetric game on F with utility function $\tilde{u}(a, a') = u_\theta(a, a', \theta)$ for $(a, a') \in F^2$. Because the game is finite and symmetric, it admits a symmetric Nash equilibrium. Let $x \in \Delta(F)$ be one; that is, for every $a \in \text{supp}(x)$,

$$\sum_{a' \in F} \tilde{u}(a, a')x[a'] \geq \sum_{a' \in F} \tilde{u}(\hat{a}, a')x[a'] \text{ for all } \hat{a} \in F. \tag{3}$$

For each $a \in F$, define

$$H(a) = \{a' \in F : (a, a') \in \mathcal{E}\},$$

and let

$$P = \sum_{(a,a') \in \mathcal{E}} x[a]x[a'].$$

We first show that $P > 0$. Suppose instead that $P = 0$. Then every pair of actions in $\text{supp}(x)$ is inefficient, so every supported action yields expected utility zero. Choose any $a \in \text{supp}(x)$. Since $a \in F$, there exists $\hat{a} \in F$ such that $(a, \hat{a}) \in \mathcal{E}$. Because $\mathcal{E}$ is symmetric, $(\hat{a}, a) \in \mathcal{E}$. Deviating to $\hat{a}$ therefore yields expected payoff at least

$$\pi(\hat{a}, a)x[a] > 0,$$

contradicting the optimality of the supported action a. Hence, $P > 0$.

Define an agreement $\tilde{\alpha}$ by conditioning the pairwise product of x on $\mathcal{E}$:

$$\tilde{\alpha}[a, a'] = \begin{cases} \dfrac{x[a]x[a']}{P}, & \text{if } (a, a') \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases}$$

By construction, $\tilde{\alpha}$ is symmetric and satisfies $\tilde{\alpha}[\mathcal{E}] = 1$.

It remains to verify the incentive constraints of a correlated equilibrium. Fix a recommendation a that occurs with positive probability and a deviation $\hat{a} \in A$. Then $x[a] > 0$. If $\hat{a} \notin F$, then $u_\theta(\hat{a}, a', \theta) = 0$ for every $a' \in A$. Because a yields strictly positive utility conditional on that recommendation, following a is strictly better than deviating to $\hat{a}$. If instead $\hat{a} \in F$, since $a \in \text{supp}(x)$, the Nash equilibrium condition (3) and the definition of $\tilde{u}(a, a')$ imply

$$\sum_{a' \in H(a)} \pi(a, a')x[a'] \geq \sum_{a' \in H(\hat{a})} \pi(\hat{a}, a')x[a'].$$

Consequently,

$$\begin{aligned} \sum_{a' \in A} \tilde{\alpha}[a, a'] \left[u_\theta(a, a', \theta) - u_\theta(\hat{a}, a', \theta)\right] &= \frac{x[a]}{P} \left[\sum_{a' \in H(a)} \pi(a, a')x[a'] - \sum_{a' \in H(a) \cap H(\hat{a})} \pi(\hat{a}, a')x[a'] \right] \\ &\geq \frac{x[a]}{P} \left[\sum_{a' \in H(a)} \pi(a, a')x[a'] - \sum_{a' \in H(\hat{a})} \pi(\hat{a}, a')x[a'] \right] \\ &\geq 0, \end{aligned}$$

where the equality comes from the definitions of $\tilde{\alpha}$ and u_θ, and the first inequality is due to the positivity of material payoffs. The above two cases together imply that for any function $g : A \to A$, we have

$$\mathbb{E}_{\tilde{\alpha}}[u_\theta(a, a', \theta)] - \mathbb{E}_{\tilde{\alpha}}[u_\theta(g(a), a', \theta)] \geq 0.$$

Since $\tilde{\alpha}$ is symmetric, the same conclusion holds for $\rho(\tilde{\alpha})$. Therefore, the incentive constraints for both sides are satisfied, meaning that $\tilde{\alpha} \in CE_{\theta,\theta}$. □

Let θ denote a morally constrained selfish type. Lemma 1 ensures that there exists a correlated equilibrium $\tilde{\alpha} \in CE_{\theta,\theta}$ such that $\tilde{\alpha}[\mathcal{E}] = 1$. Because the game between two type-θ agents is symmetric, the role-reversed agreement $\rho(\tilde{\alpha})$ also belongs to $CE_{\theta,\theta}$. Moreover, the set of correlated equilibria is defined by linear incentive constraints and is therefore convex. Hence, we have

$$\bar{\alpha} \equiv \frac{1}{2}\tilde{\alpha} + \frac{1}{2}\rho(\tilde{\alpha}) \in CE_{\theta,\theta}.$$

Since both $\tilde{\alpha}$ and $\rho(\tilde{\alpha})$ are efficient, so is $\bar{\alpha}$; moreover, the symmetrized $\bar{\alpha}$ gives both agents an expected utility of $m/2$.

Fix any competing type τ and any $\varepsilon \in (0,1)$. We claim that, in every complete-information stable outcome (μ, γ) under the population state $(\theta, \tau, \varepsilon)$, almost every type-$\theta$ agent obtains expected material payoff at least $m/2$. Suppose otherwise. Then a positive mass of type-θ agents obtain expected material payoff strictly below $m/2$. Since the utility of a morally constrained selfish agent is weakly below her material payoff, these agents also obtain expected utility strictly below $m/2$. Any two such agents can therefore rematch with one another and coordinate on $\bar{\alpha}$, which gives both agents expected utility $m/2$. This contradicts external stability. Hence, $G_\theta(\mu, \gamma) \geq m/2$.

Because each match generates at most m in total material payoff and the unit-mass population contains a total mass of $1/2$ of matches, the aggregate material payoff of the population cannot exceed $m/2$. Hence,

$$(1-\varepsilon)G_\theta(\mu,\gamma) + \varepsilon G_\tau(\mu,\gamma) \leq \frac{m}{2}.$$

Combining this bound with the preceding result that $G_\theta(\mu, \gamma) \geq m/2$ yields

$$G_\theta(\mu,\gamma) \geq \frac{m}{2} \geq G_\tau(\mu,\gamma).$$

Thus, the morally constrained selfish type θ is evolutionarily dominant.

A.1.2 Proof of Proposition 2

Let θ denote a parochial efficient type. First observe that any efficient agreement α constitutes a correlated equilibrium between two type-θ agents. Indeed, conditional on any recommendation, following it already yields the maximum joint material payoff, so no unilateral deviation can further increase utility.

Fix any competing type τ and any $\varepsilon \in (0,1)$. We first show that every complete-information stable outcome (μ, γ) under the population state $(\theta, \tau, \varepsilon)$ must exhibit perfectly assortative matching, i.e., $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$. By contradiction, suppose instead that $\mu_\theta[\tau] > 0$. Then a positive mass of type-θ agents obtain an expected utility of 0. Two such agents can rematch with one another and coordinate on an efficient agreement $\tilde{\alpha}$. Since the efficient agreement $\tilde{\alpha}$ is indeed a correlated equilibrium between two type-θ agents, which yields expected utility $\mathbb{E}_{\tilde{\alpha}}[\pi(a,a') + \pi(a',a)] = m > 0$ to both sides, external stability is violated.

We next show that almost every match between two type-θ agents exhibits efficient play. Suppose some $\alpha \in \mathrm{supp}(\gamma_{\theta,\theta})$ is inefficient. Then the type-θ agents in these matches obtain expected utility strictly below m. Any two such agents can profitably rematch and coordinate on an efficient agreement, again contradicting external stability. Hence, every $\alpha \in \mathrm{supp}(\gamma_{\theta,\theta})$ satisfies $\alpha[\mathcal{E}] = 1$.

We then have

$$\begin{aligned} G_\theta(\mu,\gamma) &= \int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a,a')]\, d\gamma_{\theta,\theta}[\alpha] \\ &= \int_{\mathcal{A}} \mathbb{E}_\alpha\left[\frac{\pi(a,a')+\pi(a',a)}{2}\right] d\gamma_{\theta,\theta}[\alpha] \\ &= \frac{m}{2}. \end{aligned}$$

The second equality follows from the exchangeability of $\gamma_{\theta,\theta}$, and the last from the efficiency of every agreement in its support.

Finally, perfect assortativity implies that type-τ agents are matched only among themselves. Since the joint material payoff cannot exceed m, we have $G_\tau(\mu,\gamma) \leq m/2$. Therefore,

$$G_\theta(\mu,\gamma) = \frac{m}{2} \geq G_\tau(\mu,\gamma).$$

Thus, the parochial efficient type θ is evolutionarily dominant.

A.1.3 Proof of Proposition 3

Fix a preference type θ, and suppose that

$$\hat{\alpha} \in \arg\max_{\alpha \in CE_{\theta,\theta}} \mathbb{E}_\alpha[u_\theta(a,a',\theta) + u_\theta(a',a,\theta)]$$

is inefficient. Let τ be a parochial efficient type, and fix any $\varepsilon \in (0,1)$. By Proposition 2, every complete-information stable outcome (μ,γ) under the population state $(\theta,\tau,\varepsilon)$ satisfies $G_\tau(\mu,\gamma) \geq G_\theta(\mu,\gamma)$.

It remains to show that the inequality is strict for some complete-information stable outcome. To this end, choose an outcome (μ,γ) as follows: let the matching profile be perfectly assortative, i.e., $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$; for type-$\theta$ matches, set $\gamma_{\theta,\theta}[\bar{\alpha}] = 1$ where $\bar{\alpha} = \hat{\alpha}/2 + \rho(\hat{\alpha})/2$; for type-$\tau$ matches, let all agents coordinate on some efficient agreement $\tilde{\alpha}$. This outcome is internally stable because the symmetry and convexity of $CE_{\theta,\theta}$ imply that $\bar{\alpha} \in CE_{\theta,\theta}$, while every efficient agreement is a correlated equilibrium between two parochial efficient agents.

We next verify external stability of (μ,γ). Type-τ agents already obtain their highest possible utility m, so they cannot participate in a blocking pair. No pair of type-θ agents can profitably rematch either. Suppose, to the contrary, that they can coordinate on some agreement $\alpha' \in CE_{\theta,\theta}$ that makes both sides strictly better off. Since every type-θ agent's current utility is $\mathbb{E}_{\hat{\alpha}}[u_\theta(a,a',\theta) + u_\theta(a',a,\theta)]/2$, a profitable deviation would require

$$\min\left\{\mathbb{E}_{\alpha'}[u_\theta(a,a',\theta)], \mathbb{E}_{\rho(\alpha')}[u_\theta(a,a',\theta)]\right\} > \frac{\mathbb{E}_{\hat{\alpha}}[u_\theta(a,a',\theta) + u_\theta(a',a,\theta)]}{2}.$$

But this implies

$$\mathbb{E}_{\alpha'}[u_\theta(a, a', \theta) + u_\theta(a', a, \theta)] > \mathbb{E}_{\hat{\alpha}}[u_\theta(a, a', \theta) + u_\theta(a', a, \theta)],$$

contradicting the choice of $\hat{\alpha}$. Hence, the constructed outcome is complete-information stable.

Finally, type-τ agents obtain average material payoff $m/2$. By contrast,

$$G_\theta(\mu, \gamma) = \frac{1}{2}\mathbb{E}_{\bar{\alpha}}[\pi(a, a') + \pi(a', a)] < \frac{m}{2},$$

where the strict inequality follows because $\bar{\alpha} = \hat{\alpha}/2 + \rho(\hat{\alpha})/2$ and $\hat{\alpha}$ is inefficient. Therefore,

$$G_\tau(\mu, \gamma) = \frac{m}{2} > G_\theta(\mu, \gamma).$$

Thus, the preference type θ is evolutionarily dominated.

A.1.4 Proof of Proposition 4

(i) Let θ denote a selfish type, and suppose there exists an efficient agreement $\tilde{\alpha} \in CE_{\theta,\theta}$. By the symmetry and convexity of $CE_{\theta,\theta}$, the agreement defined by $\bar{\alpha} \equiv \tilde{\alpha}/2 + \rho(\tilde{\alpha})/2$ is also a correlated equilibrium between two selfish agents. Moreover, the symmetrized $\bar{\alpha}$ is efficient and gives each agent an expected material payoff of $m/2$.

Fix any competing type τ, any $\varepsilon \in (0, 1)$, and any complete-information stable outcome (μ, γ) under the population state $(\theta, \tau, \varepsilon)$. We claim that almost every type-θ agent obtains an expected material payoff of at least $m/2$. To see this, suppose otherwise. Then a positive mass of type-θ agents obtain expected material payoffs strictly below $m/2$. Since the utility of selfish agents coincides with their material payoff, these agents can block the outcome by rematching with one another and coordinating on $\bar{\alpha}$, a contradiction. Therefore, we have $G_\theta(\mu, \gamma) \geq m/2$ for all complete-information stable outcomes.

Because the aggregate material payoff cannot exceed $m/2$, we have $(1-\varepsilon)G_\theta(\mu, \gamma)+\varepsilon G_\tau(\mu, \gamma) \leq m/2$. Together with $G_\theta(\mu, \gamma) \geq m/2$, this gives

$$G_\theta(\mu, \gamma) \geq \frac{m}{2} \geq G_\tau(\mu, \gamma).$$

Because τ, ε, and (μ, γ) were arbitrary, the selfish type θ is evolutionarily dominant.

Conversely, suppose no agreement in $CE_{\theta,\theta}$ is efficient. This implies that every agreement

$$\hat{\alpha} \in \arg\max_{\alpha \in CE_{\theta,\theta}} \mathbb{E}_\alpha[u_\theta(a, a', \theta) + u_\theta(a', a, \theta)]$$

is not efficient. Proposition 3 then implies that θ is evolutionarily dominated.

(ii) Let θ denote an efficient type. Suppose first that every action pair in $\mathcal{E}$ gives the two agents equal material payoffs. Fix any competing type τ, any $\varepsilon \in (0,1)$, and any complete-information stable outcome (μ, γ) under the population state $(\theta, \tau, \varepsilon)$.

We claim that almost every type-θ agent obtains expected utility m. To see this, suppose not. Then a positive mass of type-θ agents obtain expected utility strictly below m, and these agents can form a blocking pair with one another and coordinate on any efficient agreement, contradicting external stability. It follows that almost every type-θ agent interacts efficiently. Because every efficient action pair divides the material payoff equally, $G_\theta(\mu, \gamma) = m/2$. The aggregate payoff bound then implies

$$G_\theta(\mu, \gamma) = \frac{m}{2} \geq G_\tau(\mu, \gamma).$$

This means the efficient type θ is evolutionarily dominant.

For the converse, suppose that some efficient action pair gives the two agents unequal material payoffs. Choose

$$(a^*, a^\dagger) \in \arg\max_{(a,a') \in \mathcal{E}} \pi(a, a').$$

Because $\mathcal{E}$ is symmetric and contains an unequal payoff division, $\pi(a^*, a^\dagger) > m/2 > \pi(a^\dagger, a^*)$. Let τ denote the morally constrained selfish type.

Fix any $\varepsilon \in (0,1)$ and consider an arbitrary complete-information stable outcome (μ, γ) under the population state $(\theta, \tau, \varepsilon)$. The argument establishing Proposition 1 implies that type-τ agents obtain an average material payoff of at least $m/2$. The aggregate payoff bound therefore yields $G_\tau(\mu, \gamma) \geq m/2 \geq G_\theta(\mu, \gamma)$.

It remains to show that the inequality is strict in some stable outcome. Choose

$$c \in (0, \min\{\varepsilon, 1 - \varepsilon\})$$

and define the matching profile by $\mu_\theta[\tau] = c/(1-\varepsilon)$ and $\mu_\tau[\theta] = c/\varepsilon$, with the remaining probability assigned to same-type matches. The consistency condition holds as $(1-\varepsilon)\mu_\theta[\tau] = \varepsilon\mu_\tau[\theta] = c$. Let $\bar{\alpha} \in CE_{\tau,\tau}$ be a symmetric efficient agreement, whose existence follows from the argument establishing Proposition 1. Because $\bar{\alpha}$ is efficient, it also belongs to $CE_{\theta,\theta}$. Let $\gamma_{\theta,\theta}[\bar{\alpha}] = \gamma_{\tau,\tau}[\bar{\alpha}] = 1$. For cross-type matches, $\gamma_{\theta,\tau}[\hat{\alpha}] = 1$, where $\hat{\alpha}[a^\dagger, a^*] = 1$.

The outcome is internally stable. In every cross-type match, the type-θ agent obtains utility m, her highest feasible utility. The type-τ agent also finds it optimal to follow the agreement: any deviation that results in inefficient play gives her utility zero, whereas any deviation that preserves efficiency gives her a material payoff no greater than $\pi(a^*, a^\dagger)$, by the choice of $(a^*, a^\dagger)$.

To verify external stability, first observe that every type-θ agent obtains utility m and therefore cannot strictly benefit from any deviation. Hence, no blocking pair can involve a type-θ agent. Every type-τ agent obtains utility at least $m/2$: she obtains exactly $m/2$ in a same-type match and

$\pi(a^*, a^\dagger) > m/2$ in a cross-type match. Moreover, under any agreement α between two type-τ agents,

$$\mathbb{E}_\alpha[u_\tau(a, a', \tau)] + \mathbb{E}_{\rho(\alpha)}[u_\tau(a, a', \tau)] = m\alpha[\mathcal{E}] \leq m.$$

Thus, no agreement can make two type-τ agents, each of whom already obtains at least $m/2$, strictly better off. No blocking pair exists, so the constructed outcome is complete-information stable.

Because cross-type matches occur with mass $c > 0$ and $\pi(a^*, a^\dagger) > m/2 > \pi(a^\dagger, a^*)$, we have

$$G_\tau(\mu, \gamma) > \frac{m}{2} > G_\theta(\mu, \gamma).$$

Since ε was arbitrary, the efficient type θ is evolutionarily dominated.

A.2 Proofs for Section 4: Incomplete Information

In this section, we first elaborate on Example 6. Next, we establish the positive result under incomplete information, Propositions 6. Finally, we prove the negative result for the morally constrained selfish type, Proposition 5. We adopt this order because the proof of the positive result provides the central economic insights and develops a blocking argument that is also used to establish the negative result.

A.2.1 More on Example 6

We verify that the outcome $(\Lambda, p, q, \mu, \gamma)$ constructed in Example 6 is incomplete-information stable. Internal stability is immediate, so we focus on external stability. Note that a necessary condition for blocking is that both targeted types strictly benefit from the deviation when no nontargeted types participate.

First, consider a pair consisting of one type-θ agent and one type-τ agent. Against a type-θ partner, x is strictly dominant for the type-τ agent. Any viable deviating agreement must therefore prescribe x to that agent, against which the type-θ agent can obtain at most 2, her status quo utility. Hence, no such pair can block.

Second, consider two type-τ agents carrying either label. Every type-τ agent obtains utility of at least 6 in the status quo. When only the targeted agents participate, both value the joint material payoff, which cannot exceed 6. Hence, no pair of type-τ agents can block.

Finally, consider two type-θ_λ agents who target one another and propose an agreement $\hat{\alpha}$. Consider the deviation plan (λ, D, d) satisfying $D = \{\theta, \tau\}$, $d(\theta) = f$, and $d(\tau) = g$, where $g(\cdot) = x$.[32] Given this plan for her deviating partner, consider a type-θ_λ agent who follows $\hat{\alpha}$. Conditional on being recommended x, she obtains expected utility at most $4q_\lambda[\theta] < 2$. Conditional on being recommended y, the agent obtains utility at most 2, regardless of her partner's type. Her

[32] Such play is rational and strictly profitable for a type-τ_λ agent facing a type-θ partner: it yields utility 8, rather than her status quo utility of 6.

expected utility from following $\hat{\alpha}$ therefore cannot exceed 2, her status quo utility. Hence, no pair of type-θ_λ agents can block.

Therefore, no incomplete-information blocking pair exists, and the outcome is incomplete-information stable.

A.2.2 Proof of Proposition 6

Let θ denote a parochial efficient type, and fix an arbitrary preference type $\tau \in \Theta$. Consider any incomplete-information stable outcome $(\Lambda, p, q, \mu, \gamma)$. The proof proceeds through two lemmas.

Lemma 2. *In an incomplete-information stable outcome* $(\Lambda, p, q, \mu, \gamma)$*, if* $\lambda \in \Lambda$ *and* $q_\lambda[\theta] > 0$*, then* $\mu_\lambda[\lambda] = 1$*.*

Proof. Suppose, toward a contradiction, that $\mu_\lambda[\lambda'] > 0$ for some $\lambda' \neq \lambda$. By the consistency condition, $\mu_{\lambda'}[\lambda] > 0$ as well. Moreover, because distinct labels induce distinct type distributions, $q_\lambda[\theta] \neq q_{\lambda'}[\theta]$. Interchanging λ and λ' if necessary, assume that $q_\lambda[\theta] > q_{\lambda'}[\theta]$. Take any $\alpha' \in \operatorname{supp}(\gamma_{\lambda,\lambda'})$. We show that two type-$\theta_\lambda$ agents with current position (λ', α') can form an incomplete-information blocking pair by targeting one another and proposing an efficient agreement $\hat{\alpha}$.

Fix any deviation plan (λ, D, d) for one agent's prospective partner such that $\theta \in D$ and $d(\theta) = f$. Because $u_\theta(\cdot, \cdot, \tau) = 0$, for any deviating strategy $g : A \to A$,

$$\mathbb{E}_{\hat{\alpha}}[\mathbb{E}_{q_\lambda}[u_\theta(g(a), d(\cdot)(a'), \cdot) \mid D]] = \begin{cases} \mathbb{E}_{\hat{\alpha}}[u_\theta(g(a), a', \theta)]q_\lambda[\theta] & \text{if } \tau \in D, \\ \mathbb{E}_{\hat{\alpha}}[u_\theta(g(a), a', \theta)] & \text{if } \tau \notin D. \end{cases}$$

Since $\hat{\alpha}$ is efficient, following its recommendation yields utility m against a type-θ partner, the highest feasible utility. Hence, we have $f \in \arg\max_g \mathbb{E}_{\hat{\alpha}}[u_\theta(g(a), a', \theta)]$, which in turn implies

$$f \in \arg\max_g \mathbb{E}_{\hat{\alpha}}[\mathbb{E}_{q_\lambda}[u_\theta(g(a), d(\cdot)(a'), \cdot) \mid D]]$$

Moreover, the expected utility from following $\hat{\alpha}$ in the deviation satisfies

$$\begin{aligned} \mathbb{E}_{\hat{\alpha}}[\mathbb{E}_{q_\lambda}[u_\theta(a, d(\cdot)(a'), \cdot) \mid D]] &\geq \mathbb{E}_{\hat{\alpha}}[u_\theta(a, a', \theta)]q_\lambda[\theta] \\ &= mq_\lambda[\theta] \\ &> \mathbb{E}_{\alpha'}[u_\theta(a, a', \theta)]q_{\lambda'}[\theta] \\ &= \mathbb{E}_{\alpha'}[\mathbb{E}_{q_{\lambda'}}[u_\theta(a, a', \cdot)]]. \end{aligned}$$

The weak inequality above follows because $q_\lambda[\theta] \leq 1$ and $u_\theta(\cdot, \cdot, \theta) \geq 0$; the strict inequality follows from $q_\lambda[\theta] > q_{\lambda'}[\theta]$ and $m > 0$ being the maximum joint material payoff.

The same argument applies to the other targeted agent, who follows $\rho(\hat{\alpha})$. Thus, the two type-θ_λ agents form an incomplete-information blocking pair, contradicting the stability of $(\Lambda, p, q, \mu, \gamma)$. Therefore, $\mu_\lambda[\lambda'] = 0$ for every $\lambda' \neq \lambda$, meaning that $\mu_\lambda[\lambda] = 1$. □

Lemma 3. *In an incomplete-information stable outcome $(\Lambda, p, q, \mu, \gamma)$, if $\lambda \in \Lambda$ and $q_\lambda[\theta] > 0$, then any agreement $\alpha \in \operatorname{supp}(\gamma_{\lambda,\lambda})$ is efficient.*

Proof. By contradiction, suppose that some $\alpha' \in \operatorname{supp}(\gamma_{\lambda,\lambda})$ is inefficient. We show that two type-θ_λ agents with current position (λ, α') can form an incomplete-information blocking pair by targeting one another and proposing an efficient agreement $\hat{\alpha}$.

The argument in the proof of Lemma 2 implies that, under every deviation plan (λ, D, d) for the partner such that $\theta \in D$ and $d(\theta) = f$, following $\hat{\alpha}$ is optimal for the deviating type-θ_λ agent. Moreover, the expected utility from following $\hat{\alpha}$ in the deviation satisfies

$$\begin{aligned}
\mathbb{E}_{\hat{\alpha}}[\mathbb{E}_{q_\lambda}[u_\theta(a, d(\cdot)(a'), \cdot) \mid D]] &\geq \mathbb{E}_{\hat{\alpha}}[u_\theta(a, a', \theta)] q_\lambda[\theta] \\
&= m q_\lambda[\theta] \\
&> \mathbb{E}_{\alpha'}[u_\theta(a, a', \theta)] q_\lambda[\theta] \\
&= \mathbb{E}_{\alpha'}[\mathbb{E}_{q_\lambda}[u_\theta(a, a', \cdot)]],
\end{aligned}$$

where the strict inequality now follows from $q_\lambda[\theta] > 0$ and the fact that α' is inefficient.

The same argument applies to the other targeted agent, who follows $\rho(\hat{\alpha})$. Thus, the two type-θ_λ agents form an incomplete-information blocking pair, contradicting external stability. Hence, every agreement $\alpha \in \operatorname{supp}(\gamma_{\lambda,\lambda})$ must be efficient. □

The previous two lemmas imply that

$$\begin{aligned}
G_\theta(\Lambda, p, q, \mu, \gamma) &= \frac{1}{1-\varepsilon} \sum_{\lambda \in \Lambda} \left\{ \int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a, a')] \, d\gamma_{\lambda,\lambda}[\alpha] \right\} p[\lambda] q_\lambda[\theta] \\
&= \frac{1}{1-\varepsilon} \sum_{\lambda \in \Lambda} \left\{ \int_{\mathcal{A}} \mathbb{E}_\alpha \left[\frac{\pi(a, a') + \pi(a', a)}{2} \right] d\gamma_{\lambda,\lambda}[\alpha] \right\} p[\lambda] q_\lambda[\theta] \\
&= \frac{1}{1-\varepsilon} \sum_{\lambda \in \Lambda} \frac{m}{2} p[\lambda] q_\lambda[\theta] \\
&= \frac{m}{2}.
\end{aligned}$$

The first equality uses perfect assortativity, and the second comes from the exchangeability of $\gamma_{\lambda,\lambda}$. The third equality follows because, for every label λ with $q_\lambda[\theta] > 0$, every agreement in $\operatorname{supp}(\gamma_{\lambda,\lambda})$ is efficient. The final equality follows from the marginal condition.

Note that the aggregate material payoff of the population cannot exceed $m/2$, that is,

$$(1-\varepsilon) G_\theta(\Lambda, p, q, \mu, \gamma) + \varepsilon G_\tau(\Lambda, p, q, \mu, \gamma) \leq \frac{m}{2}.$$

Thus, we must have

$$G_\theta(\Lambda, p, q, \mu, \gamma) = \frac{m}{2} \geq G_\tau(\Lambda, p, q, \mu, \gamma).$$

Since the competing type τ, its population share ε, and the incomplete-information stable outcome were arbitrary, we can conclude that the parochial efficient type θ is evolutionarily dominant under incomplete information.

A.2.3 Proof of Proposition 5

Let θ denote a morally constrained selfish type. Relabeling a^* and $a^\dagger$ if necessary, assume that $\pi(a^*, a^\dagger) > \pi(a^\dagger, a^*)$, and consider a preference type τ with utility function

$$u_\tau(a, a', t) = \begin{cases} \mathbf{1}_{\{(a,a')\in\mathcal{E}\}} & \text{if } t = \tau, \\ 2 \cdot \mathbf{1}_{\{a=a^*\}} & \text{if } t \neq \tau. \end{cases}$$

Fix an arbitrary $\varepsilon \in (0, 1)$ and consider the population state $(\theta, \tau, \varepsilon)$. We show that θ is evolutionarily dominated by τ through two lemmas. Because the blocking arguments parallel those used in the proof of Proposition 6, we omit some details to avoid repetition.

Lemma 4. *$G_\tau(\Lambda, p, q, \mu, \gamma) \geq G_\theta(\Lambda, p, q, \mu, \gamma)$ for all incomplete-information stable outcomes $(\Lambda, p, q, \mu, \gamma)$.*

Proof. Fix an arbitrary incomplete-information stable outcome $(\Lambda, p, q, \mu, \gamma)$. We first record the following observation: if $q_\lambda[\theta] \leq 1/3$, almost every type-$\tau_\lambda$ agent must obtain expected utility of at least 1. Otherwise, two such agents can target one another and propose an efficient agreement $\hat{\alpha}$ satisfying $\hat{\alpha}[a^*, a^\dagger] = \hat{\alpha}[a^\dagger, a^*] = 1/2$. Following $\hat{\alpha}$ yields utility 1 whether or not the nontargeted type-θ_λ agents participate. If they participate, the only potentially profitable deviation is to choose a^* when $a^\dagger$ is recommended. Such a deviation is unprofitable because $2q_\lambda[\theta] \leq (1 - q_\lambda[\theta])$ as $q_\lambda[\theta] \leq 1/3$. The proposed deviation therefore constitutes an incomplete-information blocking pair, a contradiction.

Consider any $\lambda, \lambda' \in \Lambda$ such that $q_\lambda[\tau], q_{\lambda'}[\tau] > 0$ and $\mu_\lambda[\lambda'] > 0$. We claim that for every $\alpha \in \text{supp}(\gamma_{\lambda,\lambda'})$: (i) $\alpha[\mathcal{E}] = 1$; and (ii) if $q_\lambda[\tau] > q_{\lambda'}[\tau]$, then $\alpha[a^*, a^\dagger] \geq \alpha[a^\dagger, a^*]$.

For (i), suppose there exists $\alpha \in \text{supp}(\gamma_{\lambda,\lambda'})$ such that $\alpha[\mathcal{E}] < 1$. We first show that $q_\lambda[\theta], q_{\lambda'}[\theta] \leq 1/3$. If $q_\lambda[\theta] > 1/3$, choosing a^* gives a type-$\tau_{\lambda'}$ agent utility at least $2q_\lambda[\theta]$, whereas any other action gives utility at most $1 - q_\lambda[\theta]$. Internal stability therefore requires the agreement to recommend a^* to label-λ' agents with probability one. But $a^\dagger$ is the only action that can be optimal against a^* for type-θ_λ agents. Internal stability again implies $\alpha[a^\dagger, a^*] = 1$, a contradiction. Hence, $q_\lambda[\theta] \leq 1/3$. A symmetric argument gives $q_{\lambda'}[\theta] \leq 1/3$.

If $\alpha[\mathcal{E}] < 1$, at least one of $\alpha[a^*, a^\dagger]$ and $\alpha[a^\dagger, a^*]$ is strictly below $1/2$. Suppose $\alpha[a^*, a^\dagger] < 1/2$ without loss. We consider two cases. If $q_{\lambda'}[\theta] = 0$, then $\mathbb{E}_\alpha[\mathbb{E}_{q_{\lambda'}}[u_\tau(a, a', \cdot)]] = \alpha[\mathcal{E}] < 1$. If

$q_{\lambda'}[\theta] > 0$, the outcome $(a^*, a^\dagger)$ yields utility $1 + q_{\lambda'}[\theta]$, while every other action pair yields at most $1 - q_{\lambda'}[\theta]$. In particular, an inefficient pair in which the agent plays a^* yields $2q_{\lambda'}[\theta] \leq 1 - q_{\lambda'}[\theta]$ as $q_{\lambda'}[\theta] \leq 1/3$. Therefore,

$$\begin{aligned}\mathbb{E}_\alpha[\mathbb{E}_{q_{\lambda'}}[u_\tau(a, a', \cdot)]] &\leq \alpha[a^*, a^\dagger](1 + q_{\lambda'}[\theta]) + (1 - \alpha[a^*, a^\dagger])(1 - q_{\lambda'}[\theta]) \\ &= 1 + (2\alpha[a^*, a^\dagger] - 1)q_{\lambda'}[\theta] \\ &< 1,\end{aligned}$$

where the strict inequality comes from $\alpha[a^*, a^\dagger] < 1/2$ and $q_{\lambda'}[\theta] > 0$. Because expected utility is continuous in the agreement, this strict inequality holds in a neighborhood of α. Since $\alpha \in \text{supp}(\gamma_{\lambda,\lambda'})$, a positive mass of type-τ_λ agents obtain expected utility strictly below 1 in either case. This, together with $q_\lambda[\theta] \leq 1/3$, contradicts the preliminary observation. This proves (i).

For (ii), suppose that $q_\lambda[\tau] > q_{\lambda'}[\tau]$ and $\alpha[a^*, a^\dagger] < \alpha[a^\dagger, a^*]$. Because (i) holds and expected utility is continuous in the agreement, a positive mass of type-τ_λ agents obtain expected utility $q_{\lambda'}[\tau] + 2\alpha[a^*, a^\dagger](1 - q_{\lambda'}[\tau]) < 1$. Because $\alpha[a^\dagger, a^*] > 0$, obedience when $a^\dagger$ is recommended requires $q_{\lambda'}[\theta] \leq 1/3$. Moreover, $q_\lambda[\tau] > q_{\lambda'}[\tau]$ implies $q_\lambda[\theta] < q_{\lambda'}[\theta]$. Hence, $q_\lambda[\theta] \leq 1/3$, contradicting the preliminary observation. This proves (ii).

Next, we show that any label containing only type θ must be matched with itself. By contradiction, suppose $q_{\lambda_\theta}[\theta] = 1$ and $\mu_{\lambda_\theta}[\lambda] > 0$ for some $\lambda \neq \lambda_\theta$. Because distinct labels induce distinct type distributions, $q_\lambda[\tau] > 0$. A type-τ_λ agent facing a type-θ_{λ_θ} partner must play a^*, so the latter's current utility is at most $\pi(a^\dagger, a^*) < m/2$. Two type-$\theta_{\lambda_\theta}$ agents can instead target one another and play $\hat{\alpha}$, which gives each of them utility $m/2$. This forms an incomplete-information blocking pair. Therefore, $\mu_{\lambda_\theta}[\lambda_\theta] = 1$.

It remains to compare average material payoffs. Define

$$\Pi_{\lambda,\lambda'} = \int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a, a')]\, d\gamma_{\lambda,\lambda'}[\alpha]\,.$$

Consider any $\lambda, \lambda' \in \Lambda$ such that $\mu_\lambda[\lambda'] > 0$ and $q_\lambda[\tau] + q_{\lambda'}[\tau] > 0$. Because every pure-$\theta$ label is matched with itself, we must have $q_\lambda[\tau], q_{\lambda'}[\tau] > 0$. Claim (i) therefore implies $\Pi_{\lambda,\lambda'} + \Pi_{\lambda',\lambda} = m$. Moreover, Claim (ii) implies that the label with a higher proportion of type-τ agents obtains a weakly higher material payoff. Hence,

$$\Pi_{\lambda,\lambda'} q_\lambda[\tau] + \Pi_{\lambda',\lambda} q_{\lambda'}[\tau] \geq (\Pi_{\lambda,\lambda'} + \Pi_{\lambda',\lambda}) \frac{q_\lambda[\tau] + q_{\lambda'}[\tau]}{2} = (q_\lambda[\tau] + q_{\lambda'}[\tau]) \frac{m}{2}.$$

Therefore,

$$G_\tau(\Lambda, p, q, \mu, \gamma) = \frac{1}{2\varepsilon} \sum_{\lambda,\lambda' \in \Lambda} p[\lambda]\mu_\lambda[\lambda'](\Pi_{\lambda,\lambda'} q_\lambda[\tau] + \Pi_{\lambda',\lambda} q_{\lambda'}[\tau])$$

$$
\begin{aligned}
&\geq \frac{1}{2\varepsilon} \sum_{\lambda,\lambda' \in \Lambda} p[\lambda]\mu_\lambda[\lambda'](q_\lambda[\tau] + q_{\lambda'}[\tau])\frac{m}{2} \\
&= \frac{1}{\varepsilon} \sum_{\lambda \in \Lambda} p[\lambda] q_\lambda[\tau] \frac{m}{2} \\
&= \frac{m}{2}.
\end{aligned}
$$

The first equality uses the consistency condition, while the penultimate equality additionally uses the fact that $\sum_{\lambda' \in \Lambda} \mu_\lambda[\lambda'] = 1$. The final equality follows from the marginal condition. Because the aggregate material payoff of the population cannot exceed $m/2$, it follows that

$$
G_\theta(\Lambda, p, q, \mu, \gamma) \leq \frac{m}{2} \leq G_\tau(\Lambda, p, q, \mu, \gamma),
$$

as desired. □

Lemma 5. *$G_\tau(\Lambda, p, q, \mu, \gamma) > G_\theta(\Lambda, p, q, \mu, \gamma)$ for some incomplete-information stable outcome $(\Lambda, p, q, \mu, \gamma)$.*

Proof. Fix the population state $(\theta, \tau, \varepsilon)$ and consider an outcome $(\Lambda, p, q, \mu, \gamma)$ as follows. Choose $q_\lambda[\theta]$ such that

$$
0 < q_\lambda[\theta] \leq \frac{\pi(a^\dagger, a^*)}{\pi(a^*, a^\dagger)} \quad \text{and} \quad q_\lambda[\theta] < 2(1 - \varepsilon).
$$

There are three labels $\Lambda = \{\lambda_\theta, \lambda, \lambda_\tau\}$, where $q_{\lambda_\theta}[\theta] = q_{\lambda_\tau}[\tau] = 1$. Set $p[\lambda] = p[\lambda_\tau] = \varepsilon/(2 - q_\lambda[\theta])$, and let $\mu_{\lambda_\theta}[\lambda_\theta] = \mu_\lambda[\lambda_\tau] = \mu_{\lambda_\tau}[\lambda] = 1$. Finally, let $\gamma_{\lambda_\theta,\lambda_\theta}[\hat{\alpha}] = 1$ where $\hat{\alpha}[a^*, a^\dagger] = \hat{\alpha}[a^\dagger, a^*] = 1/2$ and $\gamma_{\lambda,\lambda_\tau}[\tilde{\alpha}] = 1$ where $\tilde{\alpha}[a^\dagger, a^*] = 1$. Figure 3 illustrates such an outcome.

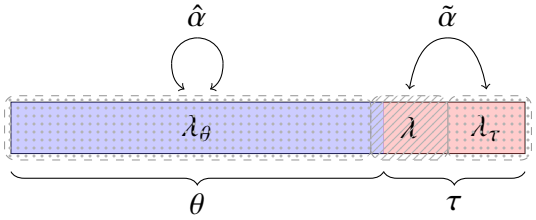

Figure 3: The outcome constructed in the proof of Lemma 5.

We argue that this outcome is incomplete-information stable. Internal stability is straightforward, so we focus on external stability. No type-θ agent can form a blocking pair with a type-τ agent. When only the targeted agents participate, the type-τ agent strictly prefers a^*, to which $a^\dagger$ is the unique optimal response for the type-θ agent. The latter therefore obtains $\pi(a^\dagger, a^*)$, which is no greater than her status quo utility. Similarly, two type-τ agents cannot form a blocking pair because their utility from interacting with one another cannot exceed 1.

It remains to consider pairs consisting of two type-θ agents. Two type-θ_{λ_θ} agents cannot block because each already obtains $m/2$, while their joint utility cannot exceed m. Now consider any pair involving a targeted type-θ_λ agent, and focus on the targeted agent whose prospective partner has label λ. For any proposed agreement α', consider the deviation plan (λ, D, d) satisfying $D = \{\theta, \tau\}$, $d(\theta) = f$, and $d(\tau) = g$, where $g(\cdot) = a^*$.[33] Given this plan for her deviating partner, the agent's expected utility from following α' can be bounded recommendation by recommendation. Conditional on being recommended a^*, she obtains at most

$$q_\lambda[\theta]\pi(a^*, a^\dagger) \leq \pi(a^\dagger, a^*).$$

Conditional on being recommended $a^\dagger$, the agent obtains at most $\pi(a^\dagger, a^*)$, regardless of the partner's type. Any other recommendation yields zero utility. It follows that her expected utility from following α' cannot exceed $\pi(a^\dagger, a^*)$. The resulting utility is no greater than the status quo utility of a type-θ_λ agent and is strictly below that of a type-θ_{λ_θ} agent. Thus, no blocking pair can involve a type-θ_λ partner.

Hence, the outcome $(\Lambda, p, q, \mu, \gamma)$ is incomplete-information stable. It is left to verify that $G_\tau(\Lambda, p, q, \mu, \gamma) > m/2 > G_\theta(\Lambda, p, q, \mu, \gamma)$. To see this, note first that play is efficient across all matches. Second, more than half of type-τ agents occupy the advantageous side of efficient play, choosing a^* against $a^\dagger$, whereas more than half of type-θ agents occupy the disadvantaged side. □

A.3 Proofs for Section 5: Polymorphism

A.3.1 Proof of Proposition 7

We reformulate our model as a roommate market in the sense of Carmona and Laohakunakorn (2024). We then show how a stable roommate matching in their framework induces a complete-information stable outcome in ours. Finally, we state and apply their existence result.

Given the primitives of our model, define a **roommate market** $\mathcal{E} = (\Theta_\nu, \nu, C, \mathbb{C}, (\succ_t)_{t\in\Theta_\nu})$ as follows. Fix an arbitrary strict total order $>$ on Θ_ν. For each $t \in \Theta_\nu$, define the set of **fair correlated equilibria** between two type-t agents by[34]

$$CE^F_{t,t} = \left\{\alpha \in CE_{t,t} : \mathbb{E}_\alpha[u_t(a, a', t)] = \mathbb{E}_{\rho(\alpha)}[u_t(a, a', t)]\right\}.$$

This set is nonempty. Indeed, for any $\alpha \in CE_{t,t}$, convexity of $CE_{t,t}$ and its invariance under role reversal imply that $\bar{\alpha} = \alpha/2 + \rho(\alpha)/2 \in CE^F_{t,t}$.

[33] Such play is rational and strictly profitable for a type-τ_λ agent paired with a type-θ agent: it yields utility 2, compared with her status quo utility of 1.

[34] The formal definition of a roommate market in Carmona and Laohakunakorn (2024) requires preferences under a given contract to be invariant to the two roles. We therefore restrict contracts between same-type agents to fair correlated equilibria. This restriction is without loss for stable outcomes: if an agreement gave the two roles different expected utilities, two agents occupying the less favorable role could rematch under its symmetrization and both strictly gain.

Let $\emptyset$ denote an unmatched partner and write $\bar{\Theta}_\nu = \Theta_\nu \cup \{\emptyset\}$. The **contract space** is $C = \mathcal{A}$. Define the **contract correspondence** $\mathbb{C} : \Theta_\nu \times \bar{\Theta}_\nu \rightrightarrows C$ by

$$\mathbb{C}(t,t') = \begin{cases} CE_{t,t'} & \text{if } t,t' \in \Theta_\nu \text{ and } t > t', \\ CE_{t',t} & \text{if } t,t' \in \Theta_\nu \text{ and } t < t', \\ CE^F_{t,t} & \text{if } t = t', \\ \mathcal{A} & \text{if } t' = \emptyset. \end{cases}$$

In particular, we have $\mathbb{C}(t,t') = \mathbb{C}(t',t)$ for all $t,t' \in \Theta_\nu$. Choose $\underline{u} < \min_{a,a',t,t'} u_t(a,a',t')$. For each type $t \in \Theta_\nu$, let $\succ_t$ be a binary relation on $\overline{\Theta}_\nu \times C$ induced by the following utility function:

$$v_t(t',\alpha) = \begin{cases} \mathbb{E}_\alpha[u_t(a,a',t')] & \text{if } t \geq t', \\ \mathbb{E}_{\rho(\alpha)}[u_t(a,a',t')] & \text{if } t < t', \\ \underline{u} & \text{if } t' = \emptyset. \end{cases}$$

By construction, these preferences only depend on the opponent's type and contract, but not on the role of the agents.

A **roommate matching** is a finite Borel measure $\varphi \in \mathcal{M}(\Theta_\nu \times \overline{\Theta}_\nu \times C)$ such that $(t,t',\alpha) \in \operatorname{supp}(\varphi)$ implies $\alpha \in \mathbb{C}(t,t')$ and, for every $t \in \Theta_\nu$,

$$\nu[t] = \varphi[\{t\} \times \Theta_\nu \times \mathcal{A}] + \varphi[\Theta_\nu \times \{t\} \times \mathcal{A}] + \varphi[\{t\} \times \{\emptyset\} \times \mathcal{A}]. \tag{4}$$

The three terms are the masses of type-t agents recorded on the first side, on the second side, and as unmatched, respectively.

For each $t \in \Theta_\nu$, define the set of type t's current positions by

$$\begin{aligned} P_t(\varphi) = &\{(t',\alpha) \in \Theta_\nu \times \mathcal{A} : (t,t',\alpha) \in \operatorname{supp}(\varphi) \text{ or } (t',t,\alpha) \in \operatorname{supp}(\varphi)\} \\ &\cup \{(\emptyset,\alpha) : (t,\emptyset,\alpha) \in \operatorname{supp}(\varphi)\}. \end{aligned}$$

The targets of a type-t agent are

$$T_t(\varphi) = \{(t',\alpha) \in \Theta_\nu \times \mathcal{A} : \alpha \in \mathbb{C}(t,t') \text{ and } \exists p' \in P_{t'}(\varphi) \text{ s.t. } (t,\alpha) \succ_{t'} p'\}.$$

Thus, $(t',\alpha) \in T_t(\varphi)$ if a type-t agent can attract a type-t' partner under contract α. Let $\bar{T}_t(\varphi) = T_t(\varphi) \cup (\{\emptyset\} \times \mathcal{A})$. A roommate matching φ is **stable** if there do not exist $t \in \Theta_\nu$, $p \in P_t(\varphi)$, and $\hat{p} \in \bar{T}_t(\varphi)$ such that $\hat{p} \succ_t p$.

The stability notions in the two frameworks are equivalent. For our existence result, however, only the following direction is needed.

Lemma 6. *Suppose there exists a stable roommate matching in the roommate market $\mathcal{E}$. Then there exists a complete-information stable outcome under population distribution ν.*

Proof. Let φ be a stable roommate matching. No unmatched position can belong to its support. Otherwise, suppose $(\emptyset, \alpha) \in P_t(\varphi)$ and choose any $\bar{\alpha} \in CE^F_{t,t}$. Because both roles under $\bar{\alpha}$ yield utility strictly above $\underline{u}$, two unmatched type-t agents could match with one another under $\bar{\alpha}$, contradicting the stability of φ.

Next, we construct an outcome (μ, γ) in our setting from the roommate matching φ. For $t, t' \in \Theta_\nu$, define the measure

$$\varphi_{t,t'}[H] = \varphi[\{t\} \times \{t'\} \times H]$$

for every measurable $H \subseteq \mathcal{A}$, and let

$$\mu_t[t'] = \frac{1}{\nu[t]}(\varphi_{t,t'}[\mathcal{A}] + \varphi_{t',t}[\mathcal{A}]).$$

The absence of unmatched agents and condition (4) imply that $\mu_t \in \Delta(\Theta_\nu)$. Moreover, $\nu[t]\mu_t[t'] = \nu[t']\mu_{t'}[t]$, so μ satisfies the consistency condition. For distinct types $t > t'$ such that $\mu_t[t'] > 0$, define

$$\gamma_{t,t'}[H] = \frac{\varphi_{t,t'}[H] + \varphi_{t',t}[H]}{\nu[t]\mu_t[t']} \quad \text{and} \quad \gamma_{t',t}[H] = \gamma_{t,t'}[\rho(H)]$$

for every measurable $H \subseteq \mathcal{A}$. For each t such that $\mu_t[t] > 0$, define

$$\gamma_{t,t}[H] = \frac{\varphi_{t,t}[H] + \varphi_{t,t}[\rho(H)]}{\nu[t]\mu_t[t]}.$$

For zero-mass pairs, choose the agreement distributions jointly so that exchangeability holds. The resulting agreement profile is therefore exchangeable.

Finally, we show that (μ, γ) is complete-information stable. First, it satisfies internal stability. This is because $(t, t', \alpha) \in \text{supp}(\varphi)$ implies $\alpha \in \mathbb{C}(t, t')$: for distinct types, every contract in the numerator is a correlated equilibrium written in the canonical orientation; for same-type matches, both $\varphi_{t,t}$ and its role reversal are supported on $CE_{t,t}$.

It remains to verify external stability. Note that every current position in the induced outcome corresponds to a current position in φ with the same expected utility. There are two cases to consider.

Suppose the induced outcome admitted a complete-information blocking pair consisting of two distinct types. Orient the proposed agreement according to the fixed ordering of types, reversing it if necessary. The resulting contract belongs to the relevant contract set and makes both agents strictly better off than their corresponding positions in φ. It therefore generates a roommate blocking pair, contradicting the stability of φ.

Suppose two agents of the same type t with current utilities u_1 and u_2 form a complete-information blocking pair through an agreement $\hat{\alpha} \in CE_{t,t}$. The symmetrized agreement $\bar{\alpha} = \hat{\alpha}/2 + \rho(\hat{\alpha})/2$

belongs to $CE^F_{t,t}$ and gives both agents

$$\mathbb{E}_{\bar{\alpha}}[u_t(a, a', t)] > \min\{u_1, u_2\}.$$

Consequently, two type-t agents drawn from the worse of the two current positions can form a roommate blocking pair under $\bar{\alpha}$, again contradicting the stability of φ. The induced outcome (μ, γ) is therefore complete-information stable. □

We say that the roommate market $\mathcal{E}$ is **acyclic** if $\succ_t$ is acyclic for each $t \in \Theta_\nu$.[35] Moreover, $\mathcal{E}$ is **continuous** if $\{(t, \alpha, t', \alpha', t^*) \in (\overline{\Theta}_\nu \times C)^2 \times \Theta_\nu : (t, \alpha) \succ_{t^*} (t', \alpha')\}$ is open, $\mathbb{C}$ is continuous with nonempty and compact values, and $\Theta_\nu \times C$ is closed. We use the following existence result, which is Corollary 2 of Carmona and Laohakunakorn (2024).

Lemma 7 (Carmona and Laohakunakorn, 2024)**.** *If $\mathcal{E}$ is an acyclic and continuous roommate market without matching externalities, then $\mathcal{E}$ admits a stable roommate matching.*

It remains to verify these conditions. Because $\succ_t$ is induced by a continuous utility function v_t, it is acyclic and its strict-preference graph is open. Moreover, the market has no matching externalities because v_t does not depend on the aggregate matching. Because A is finite, $\mathcal{A}$ is compact. Each $CE_{t,t'}$ is a nonempty compact polytope, while each $CE^F_{t,t}$ is nonempty and compact because it is the intersection of $CE_{t,t}$ with the zero set of a continuous linear function. Since Θ_ν is finite, $\mathbb{C}$ is continuous with nonempty compact values, and $\Theta_\nu \times C$ is closed. Lemmas 7 and 6 therefore establish the existence of a complete-information stable outcome under population distribution ν.

A.3.2 Proof of Proposition 8

For part (i), fix a complete-information stable outcome (μ, γ) under ν and suppose $G_\theta(\mu, \gamma) > G_\tau(\mu, \gamma)$ for some $\theta, \tau \in \Theta_\nu$. Because the aggregate material payoff of the population cannot exceed $m/2$, we must have $G_t(\mu, \gamma) < m/2$ for some $t \in \Theta_\nu$. For if not, the strict payoff inequality between θ and τ would imply that the aggregate material payoff exceeds $m/2$. Choose a parochial efficient type $t' \notin \Theta_\nu$, which is possible because Θ_ν is finite and the class of parochial efficient preferences is infinite. For some $\varepsilon \in (0, \bar{\varepsilon})$, let $\tilde{\nu} = (1 - \varepsilon)\nu + \varepsilon\delta_{t'}$. Construct an outcome $(\tilde{\mu}, \tilde{\gamma})$ under $\tilde{\nu}$ by preserving the matching and play of all incumbent types, while matching type-t' agents only with themselves and having them play efficient agreements. Because the masses of all incumbent types are scaled by the same factor, $\tilde{\mu}$ satisfies consistency. Moreover, the incumbents' current positions are unchanged, while type-t' agents attain their highest feasible utility in same-type matches. Thus, $(\tilde{\mu}, \tilde{\gamma})$ is complete-information stable. Yet, $G_{t'}(\tilde{\mu}, \tilde{\gamma}) = m/2 > G_t(\tilde{\mu}, \tilde{\gamma})$, contradicting the local evolutionary dominance of ν. This proves part (i).

[35] A relation $\succ$ on a set Z is acyclic if there is no finite sequence $\{z_1, z_2, \ldots, z_n\}$ such that $z_1 \succ z_2 \succ \cdots \succ z_n \succ z_1$.

For part (ii), suppose $\mu_t[t'] > 0$ and some $\alpha \in \operatorname{supp}(\gamma_{t,t'})$ is inefficient. Expected joint material payoff is continuous in the agreement and strictly below m at α. It is therefore strictly below m in a neighborhood of α, which has positive probability because $\alpha \in \operatorname{supp}(\gamma_{t,t'})$. Aggregate material payoff is consequently strictly below $m/2$. By part (i), all incumbent types then earn strictly less than $m/2$. The parochial efficient entrant constructed above again yields a contradiction. Hence, every agreement played in a positive-mass match is efficient. Part (i) and the aggregate payoff identity then imply that, in every complete-information stable outcome under ν, $G_t(\mu, \gamma) = m/2$ for every $t \in \Theta_\nu$.

We next show that every positive-mass cross-type match divides material payoff equally on average. For $\mu_t[t'] > 0$, let

$$h_{t,t'} = \int_{\mathcal{A}} \mathbb{E}_\alpha[\pi(a, a')] \, d\gamma_{t,t'}[\alpha] \, .$$

Efficiency and exchangeability then imply $h_{t,t'} + h_{t',t} = m$. In particular, $h_{t,t} = m/2$.

We now show that $h_{t,t'} = m/2$ even for cross-type matches as well. Suppose, toward a contradiction, that $h_{t,t'} > m/2$ for some distinct $t, t' \in \Theta_\nu$. Type t' then receives less than $m/2$ in these matches. Since its average material payoff is $m/2$, it must receive more than $m/2$ in matches with some other type. Repeating this argument and using the finiteness of Θ_ν, we obtain distinct types $t_1, \ldots, t_k$ such that $\mu_{t_i}[t_{i+1}] > 0$ and $h_{t_i,t_{i+1}} > m/2$ for every $i = 1, \ldots, k$, with indices interpreted modulo k so that $t_{k+1} = t_1$. We consider two cases.

Suppose first that k is even. For $\zeta > 0$ sufficiently small, define a new μ' by increasing the mass of matches between t_i and t_{i+1} by ζ when i is odd and decreasing it by ζ when i is even, leaving all other match masses and the agreement profile unchanged.[36] For each type, the increase in matches with one neighboring type is exactly offset by the decrease in matches with the other. Thus, μ' remains consistent and preserves every type's total mass. For sufficiently small ζ, every cycle edge remains positive. Hence, the set of current positions is unchanged, and (μ', γ) remains complete-information stable. However, the resulting outcome satisfies $G_{t_i}(\mu', \gamma) > G_{t_{i+1}}(\mu', \gamma)$ for all odd i, contradicting part (i).

Suppose next that k is odd. Relabel the cycle so that $h_{t_1,t_2} \geq h_{t_i,t_{i+1}}$ for all $i = 1, \ldots, k$. For $\varepsilon \in (0, \bar{\varepsilon})$ sufficiently small, consider $\tilde{\nu} = (1 - \varepsilon)\nu + \varepsilon\delta_{t_1}$. Scale every original match mass by $1 - \varepsilon$. Along the cycle, additionally increase the mass of matches between t_i and t_{i+1} by $\varepsilon/2$ when i is odd and decrease it by $\varepsilon/2$ when i is even.[37] Because k is odd, the masses of matches between t_1 and its two neighboring types each increase by $\varepsilon/2$, together absorbing the additional mass ε of type t_1. For every other type, the increase in matches with one neighbor is exactly offset by the decrease in

[36] More specifically, for each odd i, add $\zeta/\nu[t_i]$ to $\mu_{t_i}[t_{i+1}]$ and subtract it from $\mu_{t_i}[t_{i-1}]$; for each even i, reverse these adjustments. Leave all other components unchanged.

[37] Begin with the scaled profile $\tilde{\mu}_t[\cdot] = (1 - \varepsilon)\nu[t]\mu_t[\cdot]/\tilde{\nu}[t]$. Add $\varepsilon/(2\tilde{\nu}[t_1])$ to both $\tilde{\mu}_{t_1}[t_2]$ and $\tilde{\mu}_{t_1}[t_k]$. For each odd $i \neq 1$, add $\varepsilon/(2\tilde{\nu}[t_i])$ to $\tilde{\mu}_{t_i}[t_{i+1}]$ and subtract it from $\tilde{\mu}_{t_i}[t_{i-1}]$; for each even i, reverse these adjustments. Leave all other components unchanged.

matches with the other. The resulting matching profile $\tilde{\mu}$ is therefore consistent. For sufficiently small ε, its support is unchanged, so $(\tilde{\mu}, \gamma)$ is complete-information stable. However, it is easy to verify that $G_{t_1}(\tilde{\mu}, \gamma) \geq m/2 > G_{t_i}(\tilde{\mu}, \gamma)$ for every even i, contradicting the assumption that ν is locally evolutionarily dominant. We conclude that $h_{t,t'} = m/2$ for every positive-mass cross-type match.

Finally, fix distinct $t, t' \in \Theta_\nu$ and $\alpha \in \mathrm{supp}(\gamma_{t,t'})$. Replace $\gamma_{t,t'}$ by the point mass on α and $\gamma_{t',t}$ by the point mass on $\rho(\alpha)$, leaving all other components unchanged. The resulting outcome remains internally stable, and its set of current positions is a subset of that under (μ, γ); hence, it also remains externally stable. Applying the preceding conclusion to this outcome gives

$$\mathbb{E}_\alpha[\pi(a, a')] = \mathbb{E}_{\rho(\alpha)}[\pi(a, a')] = \frac{m}{2}.$$

This proves part (ii).