Preferences Over Games and the Evolution of Institutions

Seth Frey[1], Curtis Atkisson[2],
[1] Department of Communication, UC Davis
[2] Department of Anthropology, UC Davis

Direct correspondence to Seth Frey, sfrey@ucdavis.edu

Prepared for delivery at the Workshop on the Ostrom Workshop (WOW6) conference, Indiana University Bloomington, June 19–21, 2019.

Abstract

We introduce a framework for modeling how individuals change the games they are placed in, a process we term institutional evolution, in contrast with 'within-game' behavioral evolution. Starting at random locations in an abstract game space, agents trace trajectories through the space by repeatedly choosing between “neighboring” games, until they converge on “attractor” games that they prefer to all others. The properties of attractor games depend on the specific features that agents use to define their preferences between games. We characterize the attractors of institutional evolutionary over three types of game theoretic agent: the absolute fitness maximizing agent of economic game theory, the relative fitness maximizing agent of evolutionary game theory, and the relative group fitness maximizing agent of multi-level/group selection theory. Computing institutional change trajectories over the space of two-player ranked-outcome games, we find that the institutional evolutionary process leads to very different attractors depending on the agent. While "win-win" games account for 25% of all games in the space, this proportion is 50%, 0%, or 100% in the attractors, depending on whether agents are economic, evolutionary, or serving stable sub-groups. The first result is especially interesting: although economic agents are indifferent to the fairness of the games they choose between, the games they prefer tend incidentally to be two times more fair than baseline, as a side-effect of how preferred features co-occur. We thus present institutional evolution as a mechanism for encouraging the spontaneous emergence of cooperation among inherently selfish agents. We then investigate the sensitivity of these findings to behavioral contexts, and to games of more than two players. This work provides a flexible, testable formalism for capturing institutional evolutionary process, and for modeling the interdependencies of institutional and behavioral (between- and within- game) evolutionary processes.

Introduction

Evolutionary game theory has proven valuable for the analysis of cooperation in a wide variety of biological and social systems, and researchers in the area are increasingly characterizing the conditions necessary for fostering cooperation spontaneously. However, these results continue to treat the games agents play as fixed, disregarding the fact that agents in many game-like settings have incremental influence over the incentive structures they face, and that agents may adjust the games they play toward certain payoff structures. Across the animal world, formalisms allowing agents to change a game’s payoffs have provided parsimonious models of the intricacies of sexual selection (1), interactions with resource systems (2), and the emergence of diversity (3). And in the human world specifically, incremental changes to game structures offer a rich model of institutional change at the human level, with findings on preferences for fair punishment (4), on fairness/efficiency tradeoffs (5, 6), and negotiation processes (7). Linking within- and between- game behavior and preferences allows us to study emergent complexity in settings ranging from governance institutions to behavioral ecology.

Here we present a flexible formalism for studying the interactions of within-game "behavioral evolution" — the familiar purview of game theory — with between-game “institutional preferences” driving institutional evolution. Treating normal-form two-choice games as toy institutions, we represent institutional change dynamics in terms of trajectories over "neighboring" games, games that differ by only one payoff. By modeling institutional evolutionary dynamics explicitly, we provide a tractable formalism for generalizing beyond

static models of institution formation and testing theories of institutional change. This allows us to move beyond the simplifying assumption in institutional analysis, tacit in theories of human institutional development like Rawl's original position, that a planned institution is realized perfectly and remains stable, never deviating in practice from its design.

Background

In our framework, institutional evolution is driven by players' institutional preferences—the values and qualities people look for a social system to represent. Institutional preferences fit within a broader academic interest in human preferences between social constructions such as games, culture, norms, and language (8–11). Institutions have attracted specific interest with theories such as Binmore's, that the processes of cultural evolutionary select for institutions that are stable, efficient, and fair, in that order (12).

Research on institutional preferences and behavior over game spaces has emerged independently in several disciplines, and efforts to explore larger sets of games have become increasingly popular, both in theoretical and experimental work (13, 14). However, theoretical interest in game spaces has put more focus on static comparisons between the games in a space than dynamics across them. Previous work is primarily concerned with behavioral rather than institutional dynamics, in that it leverages large game spaces to catalog the variety of within-game dynamics (15–18), or the transfer of experience between games (19–21), rather than defining preferences over games or trajectories through them.

Among attempts to explore large game spaces, one space in particular, the Topology of Games, has attracted broad formal attempts to taxonomize or otherwise compare behavior across a range of games (22, 23) (see Fig 1a for a sample of the variety of games in the space, and Figs. 2a, S1 for 2D representations of the space). The space was first organized as a taxonomy by Rapaport (24). Its simplicity and structure make it an ideal substrate for modeling the processes of institutional evolution (25) (Fig 1b).

Institutional evolution

We introduce a framework for modeling game change as a trajectory through a space of economic games. In our formalism, agents traverse a lattice of games linked by similarity. They do so in a hill climbing process that optimizes over desirable game features such as stability, efficiency, predictability, etc. To specify this framework, we define a space, distances over it, and the elements of dynamics through it.

Institution space. In the most general terms, we understand institutional evolution as a process in which a set of people experiences an institution, makes an incremental change to it, experiences it again, and so on (Fig. 1b). The first challenge in making this picture concrete is to find a space of social systems that is rich enough to capture a range of human exchange patterns, but simple enough to remain analytically tractable.

We begin with the Topology of Games (22) a space of social systems defined in terms of the two-choice ordinal (ranked-outcome) normal-form games, an arrangement of the 144 unique ways that two agents can assign their own strict rankings over four outcomes (Figs 2a, S1). This space, including extension to n-person games in this paper, has several attractive properties. It is simple, composed of the most elementary class of economic game, and amenable to counting. It is also rich: games in the space represent a broad array of social situations (Figs. 1a, 2d). The two-player space includes many of the most famous economic games, such as Prisoner's Dilemma and Chicken, as well as social situations that have traditionally attracted little academic interest, such as no-conflict and win-win games in which individual's choices lead non-strategically to outcomes that benefit all. As mundane as these "non-game" games are, their value is clear in the fact that most of our daily social exchanges are similarly mundane. Overall, the space parsimoniously captures an impressive variety of interdependence patterns in human interactions (24, 26).

A major source of appeal for the space is its amenability to formal combinatoric approaches. There are 144 games in this space. 3/4 have exactly 1 pure strategy Nash equilibrium, 1/8 have 0, and 1/8 have two (Fib 2b). As population increases beyond 2, the number of games grows super exponentially, while the count of expected Nash equilibria per game grows more modestly, approaching a standard Poisson distribution, with a game's chances of having 0 or 1 equilibria equal to $^1/_e$=~37% and the remaining quarter have 2 or more (27, 28).

Like the number of games, the number of neighbors per game also explodes, following $(2^n)^{-(n-1)}$

This explosion in the number of neighbors per game implies an even larger explosion in the number of shortest paths between pairs of games, such that simple local strategies like hill climbing can reliably find global optima. In a sense, the increasingly convoluted topology of this space, however intriguing, probably does not meaningfully constrain trajectories through it, especially as n increases.

Distance. With the set of games in place, it is possible to introduce a simple conception of distance. We start by restricting our attention to incremental institutional change—trajectories occur over "neighboring" games. Translated to game space, two games are immediate neighbors if they differ only minimally in payoffs (Fig. 2c). Specifically, two games are neighbors if the only difference between them are the locations of a 1 and 2 ranking, a 2 and 3, or a 3 and 4. Thus, the Stag Hunt neighbors the Win-Win game because swapping the locations of a 1 payoff and a 2 payoffs will turn the former game into the latter (Fig. 1b).

A dynamic over game space. Given a space and metric over it, we can begin to specify dynamics. Agents alternate between playing the current game and selecting which neighboring game to evolve to. Thus a dynamic in this framework has three parts: the definition of player behavior within a game, the definition of a players' "institutional preferences" for selecting between a game and its neighbors, and the rules for aggregating all agents' game preferences into a single choice. Repeated as stages, a trajectory is produced by repeatedly cycling through the steps of playing a game, eliciting preferences among neighboring games, and selecting a game from the preferences produced.

One consequence of our definition of game evolution—in terms of trajectories through neighboring games in the topology—is that the resulting dynamics invite an analogy to biology's genotype/phenotype distinction: because properties such as the Nash equilibrium depend on specific patterns of payoffs across outcomes, an incremental change in a game's payoffs (its "genotype") may lead to discontinuous changes in it's strategic structure ("phenotype"). This is why the games of an otherwise uniform topology map to a discrete, discontinuous taxonomy of game types (Fig. S1).

Fig. 1b shows one possible trajectory from the Prisoner's Dilemma to a Win-Win game. A trajectory has terminated in an "attractor" game when no neighboring game is preferred to the current one.

The self-interested dynamic. Given this framework, we define a dynamic based on rational, self-interested agents who change the games they play with an eye to institutionalizing their profits and power. Where such artificially selfish agents converge on prosocial outcomes, more realistic agents are at least as likely to do the same.

Within-game, agents in this self-interested dynamic play rationally, selecting unique pure-strategy Nash equilibria when they exist, and mixed-strategy equilibria otherwise, randomizing over equilibria when several of one type exist.

Across games, a player's institutional preferences define their trajectory. Agents in the self-interested dynamic prefer games that are stable, predictable, and efficient; they prefer a game with a Nash equilibrium that is unique (stable), that is in pure strategies (predictable), and that includes the focal player's top-ranked outcome (efficient). This agent has no social

preferences: given two games that are equally stable, predictable, and efficient, players are indifferent as to which most or least benefit others.

The self-interested dynamic's aggregation rule is simple and consistent with a rational agent working to consolidate a beneficial position. The player with greater earnings after the first randomly selected game becomes the focal player choosing subsequent games to move to. Thus, power within a system confers power over it. If the first game results in tied payoffs for two players, the tie is broken randomly. Assuming that the player most recently in control of the dynamic has a small advantage, subsequent ties break in favor of the previous focal player.

After examining its attractors in three environmental contexts for the two player case, we examine the n-player case.

Two more dynamics. Demonstrating the generality of our approach, we also compute the results for two other types of game agent, those that maximize relative (rather than absolute) fitness at the individual, and those that maximize relative fitness at the group level. We represent the difference by adding social preferences to the self-interested agent, in both prosocial and anti-social varieties. Agents driving the self-interested dynamic select games based on whether the equilibrium outcome confers a maximum payoff to them. To define the relative fitness dynamic, we add an antisocial preference, defining the agent to prefer games, all else being equal, whose equilibrium maximizes the difference between their payoffs. In the group selection dynamic, the social preference is prosocial, as agents select games to maximize the sum of payoffs conferred.

Although these three types of agent—behavioral, evolutionary, and group-selected—are very different from each other, they can be united under a single conceptual umbrella. In environments that have available resources or space, where competition is low, the evolutionary regime that establishes the basis for selective pressure will select for agents who choose games to maximize personal utility (or absolute payoff, as in economic game theory) so that they may expand at the quickest rate. In environments with few free resources and no group structure, agents are instead set up to choose games to maximize relative fitness (or relative payoffs, as in evolutionary game theory). In environments with group structure and few free resources, agents choose games to maximize group fitness (by minimizing relative fitness differences, as permitted by group/multi-level selection and other related theories).

Measures

We are overall interested in the attractor games of the various dynamics and how they differ from games in the broader space. Specifically, we are interested in how inequality properties change in attractors, a question that is especially interesting in the self-interested dynamic, which does not prefer either equality or inequality. We offer two measures of equality. One is a space's proportion of "win-win" games, games in which two players share the same top-ranked outcome. Another more sensitive and continuous measure of equality is the GINI coefficient of the payoffs of each game's equilibrium outcome or outcomes.

GINI is a familiar non-parametric equality measure whose values scale between 0 (equal) and 1 (unequal), and that is easily generalizable to discrete payoffs. For an n player game, there are 2^n outcomes, each with n payoffs, distributed ranging from 1 to 2^n (the number of outcomes to rank). These may be nearly equal to each other or widely varied, a property that GINI can determine. Under this measure, an equilibrium outcome that one player ranks highly, and others rank poorly, will receive a high GINI score close to 1, while an outcome in which all players receive the same payoff (whether all high or all low) will be closer to 0, indicating high equality.

Results

Two-player games with absolute fitness maximizing agents

Our motivating questions surround the nature of institutional evolution as driven by self-interested agents. How many attractor institutions are there, how do they differ from the broader space, and, how do the values and features they represent differ from the values of the agents that selected them?

Under the self-interested dynamic, a game is an attractor if it has a unique pure-strategy Nash equilibrium that pays the maximum payoff to the focal player. The attractor games are a subset of the games with exactly one Nash equilibrium. However, not all attractors are win-win and not all win-win games are attractors. For example, the game space includes a representation of the Stag Hunt, which is win-win by our definition but has a second equilibrium that sets it outside of the set of attractors.

A feature that pops out of visualizations of the two-player space is that "win-win" games are very numerous in the two player space: one in four randomly generated two-player ordinal games are win-win (Fig. 3a).

Moving from baseline properties of the space to properties of trajectories over it, we find that 37.5% (54/144) of the two-player games are attractors, and that they form a single contiguous basin of attractors (Fig. 3b). Of games in this basin, 1/2 are win-win, compared to 1/4 of all two-player games. Thus the self-interested dynamic doubles the proportion of win-win games, despite the self-interested agent's absence of social preferences.

Two-player games with relative- and group- fitness maximizing agents

The proportional doubling of win-win games in attractor institutions holds when agents maximize personal fitness, but this result is sensitive to changes in the type of agent. Given the results of the self-interested dynamic, it is straightforward to compute the attractors of the other two. In the relative group fitness dynamic, the non-win-win games composing half of the attractors become unstable, and the proportion of win-win games becomes 100% (Fig. 3b). In the relative individual fitness dynamic, the opposite happens,, leading to a change from 25% win-win games over the whole space, to 0% win-win games in the attractor (Fig. 3b).

n-player games

All of the above results relate to two player games. Generalizing them to *n* players, we gain further insight into the effects of cross-game preferences on how institution-level selection occurs.

As the attractors of the latter two dynamics are a subset of those of the self-interested dynamic, we attend first to the properties of its attractors. We find, numerically, that attractors become a steadily decreasing proportion of games (Fig. 4a). Given these results, the results for the other two dynamics are straightforward and trivial. No additional pressure from population growth pulls the the relative fitness dynamic from 0%, nor does any pull the multi-level section group fitness dynamic from 100%. One observation, comparing across the dynamics, is that the self-interested dynamic's crashing equality makes its institutional evolutionary outcomes more difficult to distinguish from the active selection against equality that we observe in the relative fitness dynamic; socially "neutral" behavior is effectively antisocial in large-n games.

Scaling of inequality in the self-interested dynamic

Focusing again on the self-interested dynamic, we look more closely at questions of equality. Although we find that win-win games crash with *n*, their appeal as a game-level equality estimator is primarily in the ease of counting them. Alternative measures allow more sensitive judgements about the scaling of inequality in attractor institutions as the self-interested dynamic scales to larger populations.

The GINI coefficient is appealing for this task. We compare the GINI coefficients over the payoffs of Nash outcomes of attractor and non-attractor games (Fig. 4b). The difference

between them quickly becomes negligible, consistent with our other findings that, in large populations, the self-interested dynamics select for games that are only desirable to the single favored agent driving the dynamic.

Discussion

The interactions that structure our daily lives are not randomly selected from the space of social systems, nor from the small subset of systems, such as the Prisoner's Dilemma and Stag Hunt, whose prominence derives from their ability to illustrate academic points. Institutions and other social structures — languages, rites, and systems of culture — develop through a process that can be conceptualized as a trajectory through institution space. When agents have preferences over games, and the ability to make incremental changes to those games, they can dramatically remake the space of likely institutions. In particular, we find that win-win outcomes change in proportion from 25% to 0%, 50%, or 100%, depending on whether agents are motivated to maximize relative personal fitness, absolute personal fitness, or relative group fitness, respectively. The 50% case is especially illustrative: self-interested agents converge on a subset of games that are disproportionately fair in the equilibrium outcomes they provide because fairness tends to co-occur with the combination of predictability, stability, and efficiency that they seek out. Of course this particular property seems not to scale: incidental fairness very quickly becomes negligible as populations of self-interested agents grow. From the background of this result, the other two dynamics can be seen as implementing very weak "ceteris paribus" prosocial preferences.

An alternative account of norms and conventions

One contribution of this work is to offer an alternative account of informal institutional constraints like norms, conventions, and other proposed mechanisms for a state of affairs in which most daily interactions between agents in a society are rote, non-strategic, and even mundane. Existing conceptions of norms, conventions, and other proto-institutional structures understand the emergence of these constructs in terms of regularities in within-game behavior: agents can perform many actions, some combinations of actions are desirable but tenuous, agents develop a scheme for making that combination of actions likely despite strategic considerations. Under a within-game framework, the major questions are how the structure emerges and how it persists. These questions become much more straightforward under our between-game conception of institutional emergence. Our simulations support a picture in which a norm or convention emerges as a result of institutional dynamics that drive agents toward entirely “non-strategic”: win-win games in which players’ interests are naturally aligned or even no conflict “non-game” games in which their interests are orthogonal. Given a choice between a game that requires trust in another and one that does not, or one that imposes a conflict of interest and one that does not, agents will naturally select the less fraught institution. Within our institutional evolutionary framework, the convergence of a population of agents upon some stable pattern of socially efficient behavior is largely a function of institution-scale processes, rather than strictly behavioral processes. Agents who are subject to a norm or convention have not just converged on one versus another passive pattern of behavior, they actively perceive shifted payoffs, different consequences, and new strategic affordances than those who are not subject, they are *in* a different game and their collective outcome can only be modeled satisfactorily with a between-game formalism.

The pair as the most common scale of institutional organization

With general tools, we gain the ability to integrate observations from different disciplines and frameworks under a common umbrella. One classic descriptive finding in anthropology is that persistent mating pairs are an organizing principle common to many human societies. According to our findings, the stability of pairs as institutional units is due to the same

mechanisms that drive increasingly large institutions to be increasingly susceptible to unfairness and other deviations from win-win ideals. Classic findings such as the iron law of oligarchy no longer require some active theoretical mechanism: such processes are the baseline result of institutional drift, a null hypothesis against which more ideal and fragile institutions must distinguish themselves.

Effects of the evolutionary context on institutional dynamics

Each of these dynamics will arise in different population contexts. In a context of high resource availability, individuals need to be able to successfully convert resources to offspring, with little interference with conspecifics, leading to maximization of absolute payoff being selected for. As the population reaches its carrying capacity, competition with conspecifics will reward individuals who are best able to translate resources to offspring at the expense of conspecifics (who are the best competitors for jointly needed resources), leading to maximization of relative payoffs being selected for. In the high resource context, groups composed of individuals who maximize their relative payoff will perform worse than groups who maximize the total payoff of their group.

These selective regimes typify different periods of human evolution. Before the emergence of the genus *Homo*, human ancestors evolved in a relative-fitness-maximizing evolutionary regime Once members of the genus *Homo* opened a new niche of tool-assisted hunting, processing of food, and alloparenting, they transitioned from an evolutionary regime that selected on the basis of relative fitness maximization, into a new low-competition niche with pressures that instead selected for the maximization of absolute payoffs. Such a regime increases the proportion of games that are win-win and are attractors for the completely self-interested agent compared to the entire game space. Such win-win games can set the stage for the evolution of broader scale cooperation. As the human niche filled and competition increased, humans would have transitioned to the evolutionary regime that characterizes human cultural evolution: group selection that operates on the basis of relative payoff differences. Groups of agents which had settled on win-win games would have been able to out-compete other groups and single individuals. In sum, historical trajectories of completely selfish agents could have settled on win-win solutions, giving them the advantage in crowded, group-structured environments, leading to the proliferation of mutually beneficial institutions.

There are some other things I could say or we could talk about. We have earlier talked about how groups that put agents into smaller n games will be selected for because of the increased rate of mutually beneficial outcomes. It seemed like that was too complicated to include in this, though.

Limitations and future work

Our results hold for a narrow subset of games, namely those two-choice normal form games with ranked payoffs. However, extensions of this space to more choices or continuous payoffs are not likely to change our findings. Seen as equally-spaced partitions of continuous game spaces, the statistics of the ordinal games should remain proportional to those of the internal regions they define. And increasing the number of choices by two should only make our key findings stronger, particularly the decline of win-win games as population increases. More players means more outcomes, for example four outcomes for two players, eight outcomes for three players, sixteen outcomes for four players, and so forth. Thus, increasing the number of participants and outcomes inherently decreases the proportion of games that benefit all.

The topology also imposes limitations that make it impossible to test certain preferences that people are likely to have, such as a preference for extensive form games, more versus fewer choices, or more versus less information about others.

Furthermore, our agents are ultimately artificial, and establishing the generality of our findings would require comparisons to true human "institutional" preferences over games. Fortunately, where such artificially selfish agents converge on prosocial outcomes, more realistic agents are at least as likely to do the same. Against this promising background, our institutional evolutionary framework makes behavioral studies simple to articulate. In one design we have developed, two participants play a randomly selected game from the topology, and a neighboring game, and are asked to select which of the two they would like to play again. By repeating this procedure for many game pairs, an investigator can infer the games features that drive people's preferences between institutions, and directly compute the attractor games that those preferences drive dynamics towards. In fact, we explicitly developed this framework with testability in mind, making it simple to ground the assumptions of this work in behavioral data, and test its predictions about the defining features of attractor institutions.

The flexibility of the framework makes it useful for a variety of other important problems. By explicitly modeling game change dynamics, our framework makes it possible to describe other important dynamical phenomena, such as history dependence, emergent diversity, neutral evolution, the interaction of rules with culture, and the coevolution of within-game experiences and between-game preferences. For an example of possible extensions, consider the variety of aggregation rules. In the dynamics we consider here, the "winner" of a specific game outcome gains unilateral control over the next choice of game, a choice that may enable them to ensure that they continue to win. But in simple variations, the choice of game could be driven by the choice of a randomly selected agent, or, in a model of complex collective action, the preferences of a majority or plurality of players.

Conclusion

Agents in a mutable environment can change Incentive structures as those structures change them. These institutional change processes are of fundamental interest to both evolutionary and behavioral game theory in general, and institutional analysis in particular. Still, tractable frameworks for representing institutional evolution have been lacking. Dynamics over the topology of games offer a rich, tractable representation of institutional evolutionary processes. Within this framework, institutional change is understood as a trajectory over neighboring games in which players evolve the games they play in by incrementally making their payoff structures more favorable until those games reflect players' preferences for institutions. Preferences can take into account many qualities, but we focus on three simple, selfish qualities: predictability of outcomes, existence and uniqueness of Nash stable outcomes, and efficiency of outcomes. We find that the nature of "institution space" can impose constraints that encourage socially beneficial outcomes even among agents with no interest in those outcomes, but only in small social systems.

We advance a view that humans and other animals are not caged subjects of immutable institutions. Institutions evolve, often due to pressures exerted by their participants. The games we encounter are themselves the endpoints of dynamics that select and replicate certain structures over others. Elucidating the properties of "attractor" institutions sheds light on the emergence of organized human groups

Acknowledgements

The authors wish to thank Bryan Bruns, Austin Shapiro, Pete Richerson, Monique Mulder, Cristina Moya, and the EEHBC group at University of California, Davis. Author SF conceived of the research, computed the numerics, and wrote the manuscript. Author CA conceived of the

research and contributed to the manuscript. This work was supported in part by the Neukom Institute for Computational Research.

References

1. Friedman D, Magnani J, Paranjpe D, Sinervo B (2017) Evolutionary games, climate and the generation of diversity. *PLOS ONE* 12(8):e0184052.
2. Hilbe C, Šimsa Š, Chatterjee K, Nowak MA (2018) Evolution of cooperation in stochastic games. *Nature* 559(7713):246.
3. Akcay E, Roughgarden J (2011) The evolution of payoff matrices: providing incentives to cooperate. *Proc R Soc B Biol Sci* 278(1715):2198–2206.
4. Gürerk Ö, Irlenbusch B, Rockenbach B (2006) The Competitive Advantage of Sanctioning Institutions. *Science* 312(5770):108–111.
5. Mitchell G, Tetlock PE, Mellers BA (1993) Judgments of social justice: Compromises between equality and efficiency. *J Pers*.
6. Mitchell G, Tetlock PE Disentangling reasons and rationalizations: Exploring perceived fairness in hypothetical societies. eds Josh J, Kay AC, Thorisdottir H (Oxford University Press).
7. Howard N (1998) n-person 'soft' games. *J Oper Res Soc* 49(2):144–150.
8. Caldwell CA, Smith K (2012) Cultural Evolution and Perpetuation of Arbitrary Communicative Conventions in Experimental Microsocieties. *PLOS ONE* 7(8):e43807.
9. Wutich A, Brewis A, York AM, Stotts R (2013) Rules, Norms, and Injustice: A Cross-Cultural Study of Perceptions of Justice in Water Institutions. *Soc Amp Nat Resour* 26(7):795–809.
10. Heintz C, Blancke S, Scott-Phillips T (2019) Methods for studying cultural attraction. *Evol Anthropol Issues News Rev* 28(1):18–20.
11. Halim Z, Baig AR, Zafar K (2014) Evolutionary Search in the Space of Rules for Creation of New Two-Player Board Games. *Int J Artif Intell Tools* 23(02):1350028.
12. Binmore K (2005) *Natural justice* (Oxford University Press).
13. Poncela-Casasnovas J, et al. (2016) Humans display a reduced set of consistent behavioral phenotypes in dyadic games. *Sci Adv* 2(8):e1600451.
14. Moisan F, ten Brincke R, Murphy RO, Gonzalez C (2017) Not all Prisoner's Dilemma games are equal: Incentives, social preferences, and cooperation.
15. Miller JH, Page SE (2007) Complex adaptive systems: An introduction to computational models of social life.
16. Galla T, Farmer JD (2013) Complex dynamics in learning complicated games. *PNAS* 110(4):1232–1236.
17. Sanders JBT, Farmer JD, Galla T (2018) The prevalence of chaotic dynamics in games with many players. *Sci Rep* 8(4902):13.
18. Goforth D, Robinson D (2012) Effective choice in all the symmetric 2 × 2 games. *Synthese* 187(2):579–605.
19. Bednar J, Chen Y, Liu TX, Page S (2012) Behavioral spillovers and cognitive load in multiple games: An experimental study. *Games Econ Behav* 74(1):12–31.
20. Cason TN, Savikhin AC, Sheremeta RM (2012) Behavioral spillovers in coordination games. *Eur Econ Rev* 56(2):233–245.
21. Baghestanian S, Frey S (2015) GO figure: Analytic and strategic skills are separable. *J Behav Exp Econ*. doi:10.1016/j.socec.2015.06.004.
22. Robinson D, Goforth D (2005) *The Topology of the 2x2 Games* (Routledge).
23. Bruns B (2015) Names for Games: Locating 2 × 2 Games. *Games* 6(4):495–520.
24. Rapoport A (1966) A taxonomy of 2x2 games. *Gen Syst* 11:203–214.

25. Bruns B (2010) Transmuting Samaritan's Dilemmas in Irrigation Aid: An Application of the Topology of 2x2 Ordinal Games.
26. Kelley HH, et al. (2003) *An atlas of interpersonal situations* (Cambridge University Press).
27. Powers IY (1990) Limiting distributions of the number of pure strategy Nash equilibria in n-person games. *Int J Game Theory* 19(3):277–286.
28. Rinott Y, Scarsini M (2000) On the Number of Pure Strategy Nash Equilibria in Random Games. *Games Econ Behav* 33(2):274–293.

FIGURES & PLOTS

FIGURES

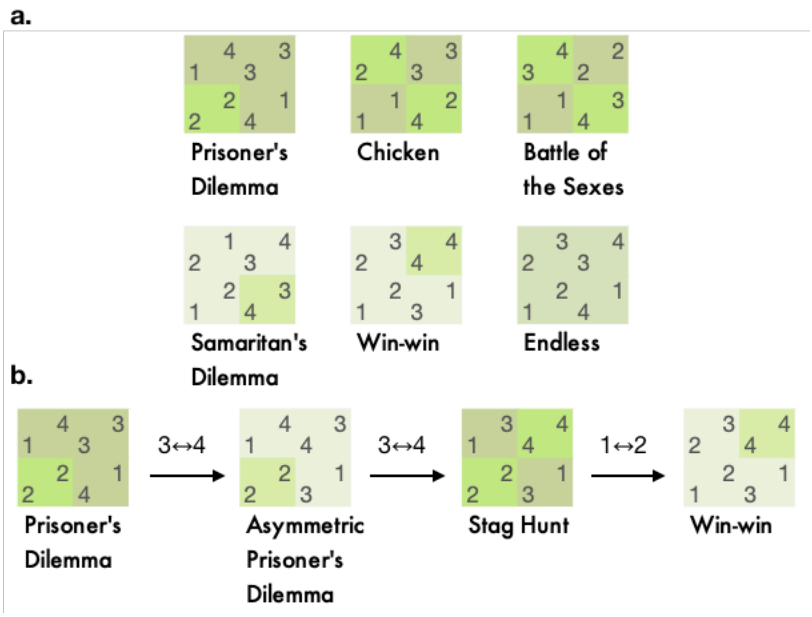

Fig 1. A sampling of ordinal games, with a trajectory through them.

a. The space of 2-player, 2-choice ordinal games includes a variety of interesting and relevant games, including well-studied games such as the Prisoner's Dilemma, Chicken, and Battle of the Sexes, and less remarked upon games such as win-win games that don't require strategy, cyclic games without pure-strategy equilibria, and asymmetric games. Outcomes that are Nash equilibria are slightly brighter. By ordinal, we mean games with consecutive integer payoffs up to the number of outcomes. In these illustrations, 4 is high and 1 is low. Ordinal games have the advantage of being amenable to counting.

b. Two games in this space are neighbors if they differ by a swap of similar payoffs. Assuming that most institutional change is incremental, institutional evolution can be modeled as a trajectory through the neighboring games. Here we illustrate how an agent might incrementally evolve a prisoner's dilemma into a win-win game. This trajectory terminates on the game "Win-win", which is an attractor for self-interested agents who prefer stable, predictable, and efficient games (defined herein as offering a unique pure-strategy Nash equilbrium that confers a maximum payoff). Example adapted from (25).

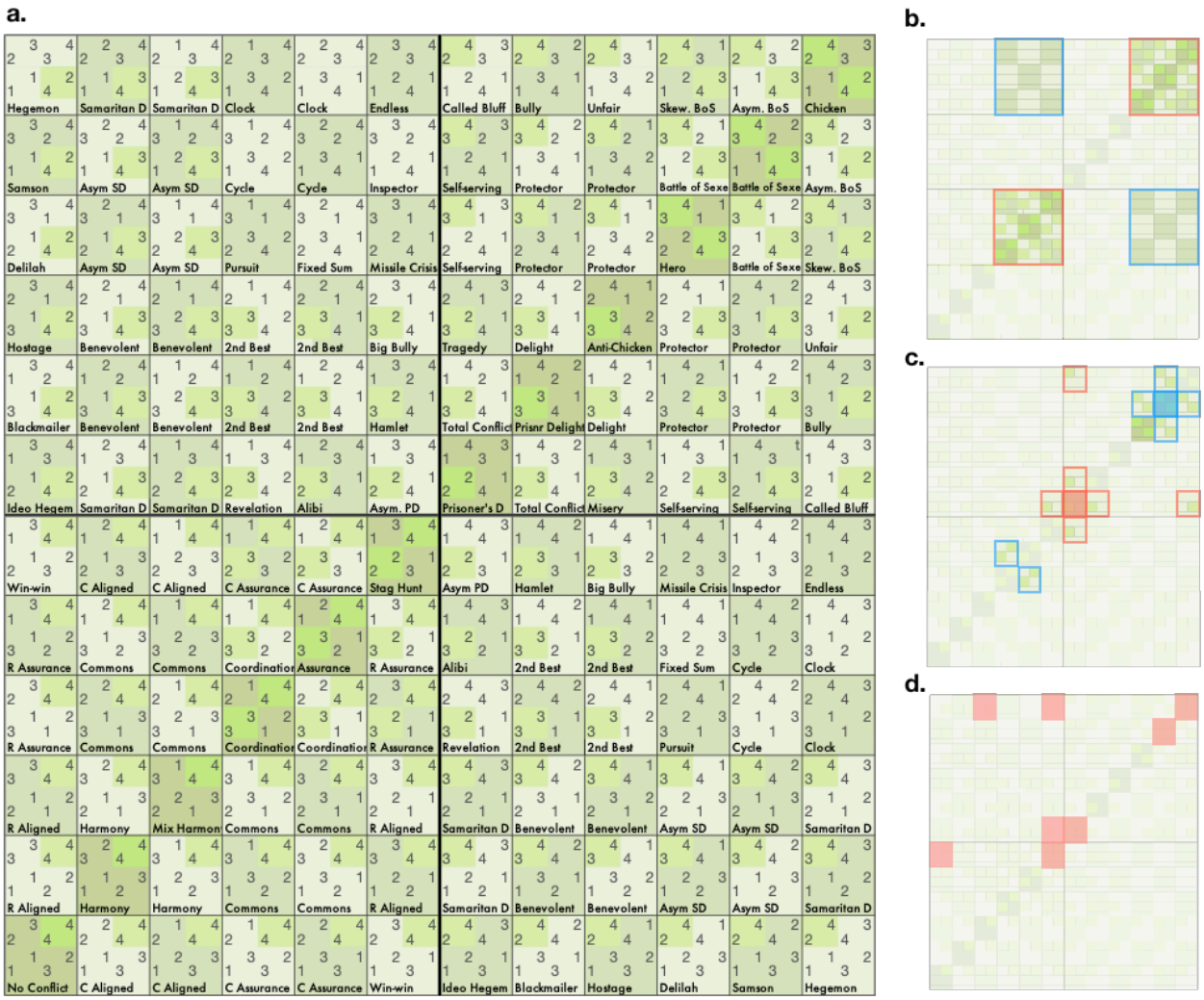

Fig 2 The space of two-player games, with masks illustrating some of its properties.

a. A simple representation of the space of 144 two-player, two-choice games with ordinally ranked payoffs. Observe that symmetric games, occupying the increasing diagonal, are a minority, and that the lower-left quarter of games are win-win, in the sense of having one outcome that confers the maximum payoff of 4 to both players (also see Fig 3a). Spaces with more than two players are much larger and more difficult to diagram than this.

b. This mask illustrates the Nash properties of games in this space. The games in the blue outlines have no pure strategy Nash equilibria. The games in red outlines have two. The remaining games, 75%, have exactly one pure strategy Nash equilibrium. We discuss how this distribution changes as the number of players increases.

c. This mask illustrates the complex nature of neighbor relations in the two-player space. The red outlines show the six "neighbors" of the prisoner's dilemma. The blue outlines show the neighbors of the Battle of the Sexes. Note that some adjacent games are not visually adjacent, a shortcoming of the grid representation of what is in truth a much more complex topology.

d. This mask of panel **a** shows the locations of the games in Figs. 1a and 1b.

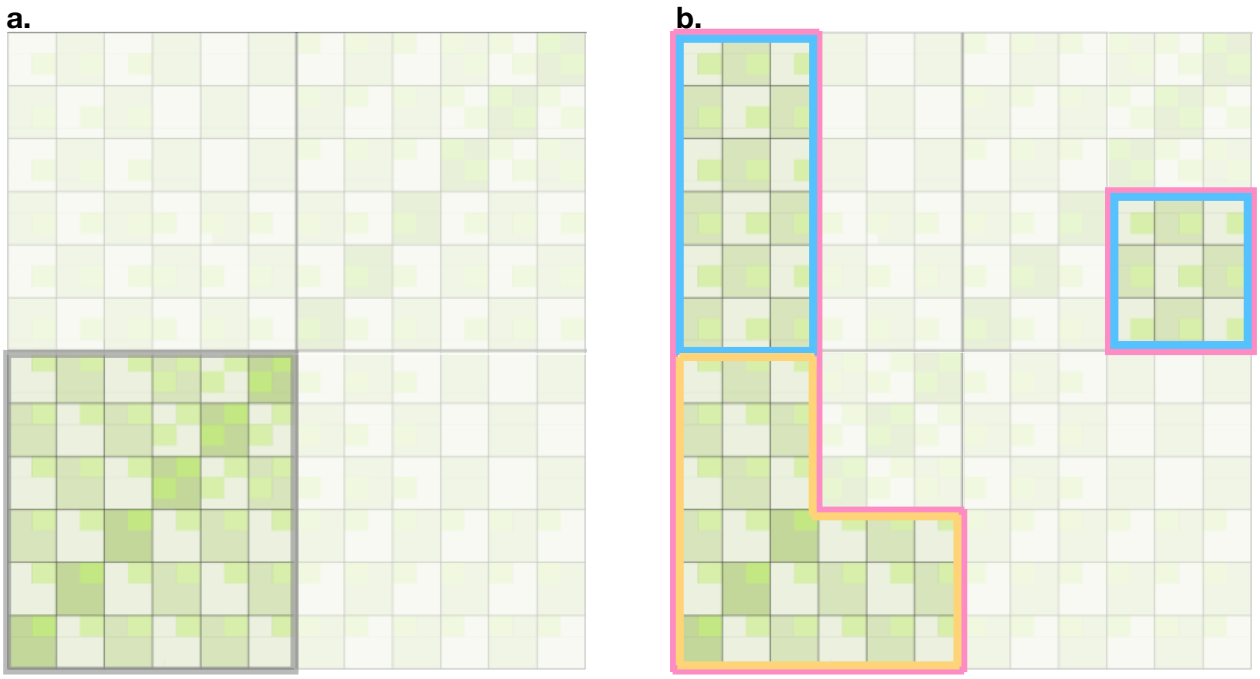

Fig 3. The win-win games and the attractor games of three institutional dynamics.

a. Win-win games account for 1/4 of the games in the two-player space, shown here in the lower-left quadrant. See Fig 2a for details of each, namely that each has an outcome conferring the maximum payoff of 4 to both players.

b. The encompassing pink outlines shows the basin of attractors that results from self-interested agents' evolutionary trajectories. Note that these attractor games form a contiguous block: as the 9 games on the right have several neighbors among the games in the block on the left, via swaps that are not apparent from this grid representation. Note also that half of these attractor games are in the win-win quadrant. Compared to panel a, the institutional evolutionary process double the chances of converging upon a win-win game, even though the selfish agents driving it have no explicit preferences for mutually beneficial games. With the attractors of the multilevel selection dynamic (orange outline), that probability increases to 100%, as all of its attractor games are win-win. Conversely, the fitness maximizing evolutionary agent will converge on the more unfair games in the upper half (green outline), all of which have an asymmetry between the focal player's earnings and those of the other player.

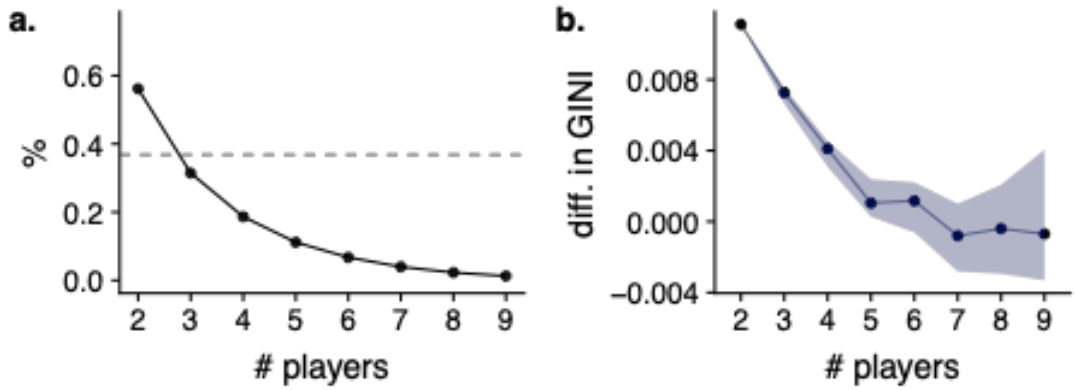

Fig 4. Emergent fairness driven by the self-interested dynamic disappears as player populations grow.

a. Attractors become a smaller fraction of games with increasing population. The black line, derived from simulation, gives the computed proportion of games up to 9 players that are attractors of the self-interested dynamic. **b.** Each game outcome contains payoffs for each player, and those payoffs can differ widely from each other. We compute the GINI coefficient of the payoffs in Nash outcomes, and compare them within the attractors and in the full space. We find that the difference quickly becomes negligible, telling a complementary story to that of Fig. 3b, that the self-interested dynamic biases institutional evolutionary processes to emergently select fair games when populations are small, an effect that disappears for larger populations.

SUPPORTING FIGURES

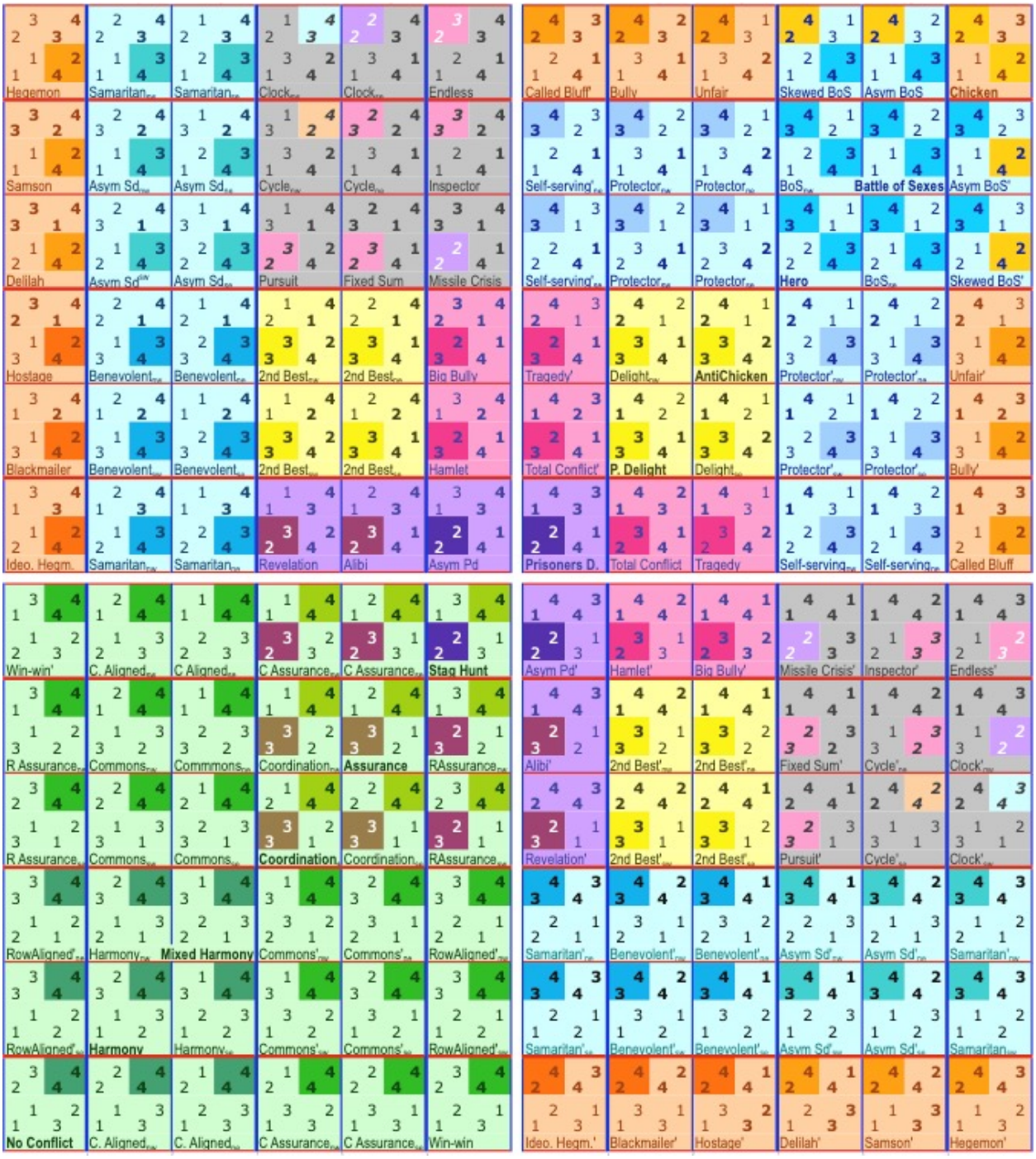

Fig S1. Richer visual representation of the topology of 2-player, 2-choice ordinal games.

This figure, an elaboration of Fig. 2a, illustrates neighboring relations in the blue, red, and quadrant boundaries, and game classes by background color. Copied with permission from (23)