The fitness landscape of social norms in social dilemmas

Maximilian Puelma Touzel[*]
Department of Computer Science
Université de Montréal
Montréal, Canada
`puelmatm@mila.quebec`

Abstract

By specifying behaviour across multiple agents, social norms are a coordination approach to resolving social dilemmas. Decentralized and wide adoption can be achieved by norms whose prescription involves interpreting stochastic signals in the environment. Such signals must have enough *correlation* to orchestrate mutually beneficial coordination and enough disincentivizing *uncertainty* about the benefits of exploiting that coordination. Evolutionary game theory of matrix games has been used to describe how, by rational agents comparing and adopting norms, a norm can evolve to become dominant in a population. Morsky and Akçay [2019] classify norms according to a set of rationality criteria and prove that norms that are consistent with reward-optimal choice dynamics of local social interactions implement correlated, rather than Nash game theoretic equilibria. Here, we present a version of this theory that clarifies the basic ingredients. We formulate it in the more general Markov game setting more commonly used in reinforcement learning theory. We illustrate the theory by mapping norms over the signal and reward space, while also giving a detailed exposition of the underlying mechanics of the approach. Finally, we give a generic analysis of replicator dynamics, which Morsky and Akçay [2019] propose as a means by which these norms could emerge.

Introduction

In the epistemic perspective on game theory [Aumann, 1987], uncertainty about opponents' strategies is framed as an inference problem that must tackled in pursuit of an optimal strategy[2]. Through any consistent calibration process [Foster and Vohra, 1997, Hart and Mas-Colell, 2000], each player learns to infer a posterior over the plays of the opponent given the player's observation of the state. The natural Bayes optimal strategy is then for a player to take the action that gives the highest posterior-averaged utility. Surprisingly, this fundamental process does not lead to the celebrated solution concept of a Nash equilibrium[3], but rather to the lesser known concept of a *correlated equilibrium*.

Morsky and Akçay [2019] present a theory for how correlated equilibria emerge endogenously through an evolution of social norms. Here, we present a minimal version of their theory (*e.g.* without partitions) using the notation of reinforcement learning theory and illustrate it with an application to the game of chicken. To demonstrate the theory, Morsky and Akçay [2019] take the well-established approach to gaining theoretical traction in evolutionary dynamics of game-playing agents by considering a dynamics wherein pairs of agents are selected at random for interaction[4].

---

[*]mptouzel.github.io

[2]Note that optimal strategies for perfect knowledge settings (*e.g.* of other agents) are always pure strategies.

[3]Which assumes, unrealistically, that the player knows the policies of the opponent.

[4]This strong assumption can be relaxed by moving to a finite population description in which interactions are specified (e.g. assortative or via a given social graph) at the expense of having to resort to numerical analysis of solutions [Hilbe, 2011].

Under this assumption, a norm selection dynamics can be faithfully represented through a particularly simple mean-field description called replicator dynamics. After presenting our exposition of the theory, we give a complete characterization of the game family of the game of chicken and then go on to give a general formulation and analysis of replicator dynamics.

Related Work

This work translates ideas at the intersection of evolutionary game theory and sociology to the multi-agent reinforcement learning research field. The concept of correlated equilibria was introduced by the works of Aumann [1987]. Sociologists then proposed correlated equilibrium as a concept central to social norms [Gintis, 2010] and game theoretic work on this topic, albeit sparse, extends back decades[Axelrod, 1986] (see Morsky and Akçay [2019]'s account of this literature). In parallel, correlated equilibria have been studied for decades in the multi-agent reinforcement learning (MARL) theory community, with seminal contributions from the group of Amy Greenwald, *e.g.* Greenwald and Hall [2003]. The recent work in that field is expanding its realm of applicability, e.g. to extensive form games Anagnostides et al. [2022] (and references therein). Separately and much more recently, the emergence of social norms has been studied empirically using deep reinforcement learning agents, e.g. Köster et al. [2022] and Vinitsky et al. [2023]. Our work aims to motivate the intermingling of these empirical and theoretical traditions by developing the evolutionary game theory of social norms as studied in evolution and social science in the setting most relevant to models of multi-deep reinforcement learning agent systems that are currently being used to explore the social dimensions of artificial intelligence.

Background

Markov Game Social Dilemmas   We focus on the two-player partially observable Markov game setting. In this simple illustration, we assume state transitions do not depend on the actions and rewards do not depend on the state. We also restrict ourselves to the symmetric case. Accordingly, we denote the player with unprimed variables, while those of the opponent are primed. All sets are discrete.

The environment is a set of states $\mathcal{S}$. State information available to agents is bounded by (possibly stochastic) observation functions $p(o|s)$ and $p(o'|s)$ for observations o and o' from observations spaces $\mathcal{O}$ and $\mathcal{O}'$ for player and opponent, respectively. Action-independent state transitions are given by $\mathcal{T}(\tilde{s}|s)$ where $\tilde{s}$ is the next state. We assume a stationary environment so the state description reduces to the stationary state distribution $p(s)$ defined implicitly by $p(\tilde{s}) = \sum_s \mathcal{T}(\tilde{s}|s)p(s)$. The (in general state-correlated) joint observation distribution is then $p(o, o') = \sum_s p(o|s)p(o'|s)p(s)$. For $\boldsymbol{p}_s$ denoting the vector representation of $p(s)$ and $\boldsymbol{P}_{oo'}$ denoting the matrix representation of $p(o, o')$, we have the matrix expression $\boldsymbol{P}_{oo'} = \boldsymbol{P}_{o|s}\mathrm{diag}(\boldsymbol{p}_s)\boldsymbol{P}_{o'|s}^\top$, where $\boldsymbol{P}_{o|s}$ is the $|\mathcal{O}| \times |\mathcal{S}|$ matrix representation of $p(o|s)$ and similarly for $p(o'|s)$[5]. With the state variable marginalized, we hereon represent the environment by $\boldsymbol{P}_{oo'}$.

A player policy (a possibly mixed strategy) $\pi : \mathcal{O} \to \Delta^{|\mathcal{A}|-1}$ is a mapping from the observation space to distributions on an action space, $\mathcal{A}$. π can be represented as a $|\mathcal{A}| \times |\mathcal{O}|$ matrix $\boldsymbol{P}_{a|o}$, such that $\sum_i (\boldsymbol{P}_{a|o})_{ij} = 1$ for all j. A deterministic policy (pure strategy) corresponds to policies for which $(\boldsymbol{P}_{a|o})_{ij} \in \{0, 1\}$. The opponent's strategy π' is defined similarly.

In practise, $\mathcal{O}$ will be a large space of which only a small part of which will be useful as signal by the agent for any given task. Here, we will assume that the agent has performed this selection and that $\mathcal{O}$ is the signal space it attends to get suggestions for how to act. Thus, we assume $|\mathcal{O}| = |\mathcal{A}|$. To

---

[5]In this two-player, stochastic policy setting, the matrix algebra of the probability theory of two-dimensional, discrete random variables is useful. Namely, for a matrix representation of a joint distribution on x and y, $\boldsymbol{P}_{xy}$, the marginals are computed as $\boldsymbol{p}_x = \boldsymbol{P}_{xy}\mathbf{1}$ and $\boldsymbol{p}_y = \boldsymbol{P}_{xy}^\top\mathbf{1}$. Independence is the factorization condition $\boldsymbol{P}_{xy} = \boldsymbol{p}_x\boldsymbol{p}_y^\top$ (when x and y are binary, $\boldsymbol{P}_{xy}$ is determined by the marginals and the correlation, ρ, $\boldsymbol{P}_{xy} = \boldsymbol{p}_x\boldsymbol{p}_y^\top + \rho\sqrt{\sigma_x^2\sigma_y^2}\left[\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}\right]$, where $\sigma_i^2 = \det(\mathrm{diag}(\boldsymbol{p}_i))$ for $i = x, y$ is the variance of x and y respectively). The expectation of a scalar function $f(x, y)$ is computed as $\mathbb{E}_{(x,y)}[f(x, y)] = \mathrm{tr}(\boldsymbol{F}\boldsymbol{P}_{xy}^\top)$, where variable $\boldsymbol{F}$ is the matrix of $f(x, y)$ function evaluations. Adding conditional dependence via a correlated pair of random variables $(\tilde{x}, \tilde{y})$ with joint distribution $\boldsymbol{P}_{\tilde{x}\tilde{y}}$ gives $\boldsymbol{P}_{xy} = \boldsymbol{P}_{y|\tilde{y}}\boldsymbol{P}_{\tilde{x}\tilde{y}}^\top\boldsymbol{P}_{x|\tilde{x}}^\top$, where $\boldsymbol{P}_{y|\tilde{y}}$ and $\boldsymbol{P}_{x|\tilde{x}}$ are the matrix representations of the respective conditional distributions.

map onto the simply symmetric matrix games discussed below, we identify the two action spaces $\mathcal{A} = \mathcal{A}'$ (and thus also the observation spaces $\mathcal{O} = \mathcal{O}'$). We also assume the observation spaces confer no player bias so the marginals $p(o)$ and $p(o')$ are equal (achieved by the symmetry constraint, $\boldsymbol{P}_{oo'} = \boldsymbol{P}_{oo'}^{\top}$).

A reward function in this symmetric player setting is a mapping $r : \mathcal{A} \times \mathcal{A}' \to \mathbb{R}$ returning the reward to a player for playing action $a_i \in \mathcal{A}$ when its opponent plays $a'_j \in \mathcal{A}'$. We can represent r using the $|\mathcal{A}| \times |\mathcal{A}|$ action reward matrix $\boldsymbol{R}$ where $r(a = a_i, a' = a_j) = R_{ij}$. Note that unlike a general Markov game, the reward here does not depend on the environment observations.

Canonical social dilemmas are two-player matrix games with a binary action space corresponding to a cooperative and exploitative action. The reward matrix for such dilemmas is

$$\boldsymbol{R} = \begin{bmatrix} B & S \\ T & P \end{bmatrix} \text{ with } \begin{array}{l} B, \text{ the benefit for mutual cooperation;} \\ P, \text{ the punishment for mutual exploitation;} \\ S, \text{ the sucker's payoff for cooperating when the other has exploited; \&} \\ T, \text{ the tempting reward for exploiting when the other has cooperated.} \end{array} \quad (1)$$

In this setting, the conditions for a tension between individual and collective interests that characterize a social dilemma [Macy and Flache, 2002] are that

1. $B > P$: mutual cooperation > mutual exploitation
2. $B > S$: mutual cooperation > being suckered
3. $2B > T + S$: mutual cooperation > equal mix of unilateral cooperation and exploitation
4. a) $T > B$ (*greed*): exploiting a cooperator > mutual cooperation
   and/or
   b) $P > S$ (*fear*): mutual exploitation > being exploited.

Condition 3 ensures cooperation offers more than pure competition (*i.e.* to exploit and be exploited). Condition 4 gives 3 classes of dilemmas

	not *greed*	*greed*
not *fear*	no dilemma $B > T > S > P$	game of chicken $T > B > S > P$
fear	stag hunt $B > T > P > S$	prisoner's dilemma $T > B > P > S$

These conditions have been applied to the sequential decision-making setting relevant to reinforcement learning to try to leverage the insight they have offered into matrix games [Leibo et al., 2017]. We will focus on the game of chicken since it serves as the canonical example of correlated equilibria.

Game of chicken example In the case of *greed* and not *fear*, $T > B > S > P$ characterizes the canonical family of anti-coordination social dilemmas known as the *game of chicken* where in one interpretation, $\mathcal{A} = \{stop, go\}$ defines the actions available to each of two cars approaching an intersection from different roads. This game family captures the inevitability of conflict in the pursuit of self-interest as a coordination dilemma: highest reward for *go*, but only when the other chooses to *stop*, otherwise the reward is low for both agents (because they crash into eachother). The car interpretation evokes the coordination solutions (traffic lights, the *yield-to-the-right* rule, etc.) that have evolved to facilitate the resolution of this dilemma. For this reason, the game is often invoked as a simple setting to explain a correlated equilibrium, a solution concept by which social dilemma are rationally resolved. In this case, the game environment is augmented with observation spaces ($\mathcal{O} = \mathcal{O}' = \{\bullet, \bullet\}$ in the traffic light solution) that provides signals to be interpreted as suggested actions (e.g. the signal-following norm: to stop when $\bullet$ is observed, and go when $\bullet$ is observed). There are multiple ways to represent this simple matrix game in the more general Markov game setting outlined above. Here, we represent the game by defining observation functions as partitions on the state space such that $(\boldsymbol{P}_{o|s})_{ij}, (\boldsymbol{P}_{o'|s})_{ij} \in \{0, 1\}$. The state space then naturally partitions into the refined partition obtained by intersecting the two player's partitions. This refined partition abstracts $\mathcal{S}$ into $|\mathcal{A}|^2 = 4$ states, $\{(\bullet, \bullet), (\bullet, \bullet), (\bullet, \bullet), (\bullet, \bullet)\}$. In a minimal example, $|\mathcal{S}| = 4$ and partitions can be chosen such that the elements of $\boldsymbol{P}_{oo'}$ are those of $\boldsymbol{p}_s = (p_1, p_2, p_3, p_4)$. For example,

$$\boldsymbol{P}_{o|s} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} , \quad \boldsymbol{P}_{o'|s} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \implies \boldsymbol{P}_{oo'} = \begin{bmatrix} p_{\bullet\bullet} & p_{\bullet\bullet} \\ p_{\bullet\bullet} & p_{\bullet\bullet} \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix} . \quad (2)$$

The symmetry assumption on $\boldsymbol{P}_{oo'}$, $p_{\bullet \bullet} = p_{\bullet \bullet}$, demands that $p_2 = p_3$. Having control over $\boldsymbol{P}_{oo'}$ here means having control over $\boldsymbol{p}_s$. However, for the general case of a large $\mathcal{S}$, given a fixed distribution $\boldsymbol{p}_s$, control over $\boldsymbol{P}_{oo'}$ could also be achieved via the freedom in partitioning through the joint sculpting of $\boldsymbol{P}_{o|s}$ and $\boldsymbol{P}_{o'|s}$.

Correlated equilibrium Let us denote by S the system without the player's policy, $S = (\pi', p(o, o'), r) = (\boldsymbol{P}_{a'|o'}, \boldsymbol{P}_{oo'}, \boldsymbol{R})$. The *average reward* ρ for the player is then

$$\rho(\pi, S) = \mathbb{E}[r(a, a')|\pi, S] := \sum_{o,o',a,a'} r(a, a')\pi'(a'|o')p(o, o')\pi(a|o) \tag{3}$$

$$= \operatorname{tr}\left(\boldsymbol{R}\boldsymbol{P}_{a'|o'}\boldsymbol{P}_{oo'}\boldsymbol{P}_{a|o}^{\top}\right), \tag{4}$$

where we omit the transpose on $\boldsymbol{P}_{oo'}$ since it is symmetric. $\rho(\pi', S')$ is defined similarly where p' is p with swapped arguments ($p' = p$ for the unbiased observation distributions we consider here) and $r' = r$ for this symmetric player setting. A policy $\boldsymbol{P}_{a|o}$ that maximizes ρ, *i.e.* $\operatorname{tr}(\boldsymbol{R}\boldsymbol{P}_{a'|o'}\boldsymbol{P}_{oo'}(\boldsymbol{P}_{a|o} - \tilde{\boldsymbol{P}}_{a|o})^{\top}) \geq 0$ for any other policy $\tilde{\boldsymbol{P}}_{a|o}$, is best in expectation over all randomness including over the observation, o, observed by the player. A stronger notion of optimality is that $\boldsymbol{P}$ is optimal for *each* observation value $o = o_i$, $((\boldsymbol{R}\boldsymbol{P}_{a'|o'}\boldsymbol{P}_{oo'})_{\cdot i})^{\top}(\boldsymbol{P}_{a|o} - \tilde{\boldsymbol{P}}_{a|o})_{\cdot i} \geq 0$ for all i[6]. Such a policy is called a *best response* (BR) policy.

More generally, we can think of a (not necessarily factorizable) joint action distribution $p(a, a') = \boldsymbol{P}_{a|o}\boldsymbol{P}_{oo'}\boldsymbol{P}_{a'|o'}^{\top}$ and the environmental observations simply as a means to generate action-action correlations between the player and its opponent. In this context, a *correlated equilibrium* given r is the observation ensemble p and a pair of deterministic policies (π, π') that are best responses to the expected behaviour of the other given their observations, *i.e.* for $\tilde{\pi}$ and $\tilde{\pi}'$ denoting any other strategy for player and opponent, respectively,

$$\begin{aligned}((\boldsymbol{R}\boldsymbol{P}_{a'|o'}\boldsymbol{P}_{oo'})_{\cdot i})^{\top}(\boldsymbol{P}_{a|o} - \tilde{\boldsymbol{P}}_{a|o})_{\cdot i} \geq 0 \quad &\text{from the player's perspective } S \text{, and} \\ ((\boldsymbol{R}\boldsymbol{P}_{a|o}\boldsymbol{P}_{oo'})_{\cdot i})^{\top}(\boldsymbol{P}_{a'|o'} - \tilde{\boldsymbol{P}}_{a'|o'})_{\cdot i} \geq 0 \quad &\text{from the opponent's perspective } S' \text{.}\end{aligned} \tag{5}$$

A *Nash equilibrium* is the special case where the joint action distribution factorizes, $p(a, a') = p(a)p(a')$, *i.e.* the factorization condition $\boldsymbol{P}_{a|o}\boldsymbol{P}_{oo'}\boldsymbol{P}_{a'|o'}^{\top} = \boldsymbol{p}_a\boldsymbol{p}_{a'}^{\top}$ into the outer product of its marginal distribution vectors.

Social norm theory

A social norm is a belief to which an agent can subscribe that characterizes a social interaction involving multiple agents. For a pair-agent interaction, a social norm for a player both *prescribes* to that player how they should act in that interaction (via the prescribed policy, $\pi_{\text{prescribed}}$, here represented by $\boldsymbol{P}_{a|o}$), and *describes* to that player how the opponent should act in the same interaction (the described policy $\pi_{\text{described}}$, here represented by a policy matrix denoted $\boldsymbol{D}_{a'|o'}$). A norm as a belief within a formal reasoning framework is thus represented as the matrix pair $(\boldsymbol{P}_{a|o}, \boldsymbol{D}_{a'|o'})$. These are deterministic assignments so $\boldsymbol{P}_{a|o}$ and $\boldsymbol{D}_{a'|o'}$ are binary-valued matrices with a single non-zero entry per column. The total number of norms is then equal to the total number of unique ordered pairs of policies, $N = (|\mathcal{A}||\mathcal{O}|)^2$ norms.

Let $n, n' \in \{1, \ldots, N\}$ index these pairs so that the average reward $\Gamma_{nn'}$ (hereon called the norm payoff) for a player playing within a norm n that indexes $(\boldsymbol{P}_{a|o}, \boldsymbol{D}_{a'|o'})$ against an opponent playing within the norm n' that indexes $(\boldsymbol{P}'_{a|o}, \boldsymbol{D}'_{a'|o'})$ is just $\rho(\pi, S)$ (eq. (4)) with $\boldsymbol{P}_{a'|o'} = \boldsymbol{P}'_{a|o}$. The norm payoff matrix is then denoted $\boldsymbol{\Gamma} = (\Gamma_{nn'})$. While the descriptive part of a norm $\boldsymbol{D}_{a'|o'}$ does not affect its payoff directly, it does affect whether it is rational for a player to adopt the norm. Morsky and Akçay [2019] provide a classification of norms via increasingly strong notions of validity (*c.f.* fig. 1). Namely, *rational* norms are those that are internally valid: the prescription $\boldsymbol{P}_{a|o}$ is never worse than any other prescription $\tilde{\boldsymbol{P}}_{a|o}$ given the description, $\boldsymbol{D}_{a'|o'}$, else they are called *null* norms. Null norms are never optimal so in that case the player plays a (observation-independent) default strategy instead (Morsky and Akçay [2019]'s choice being that the player eschews null norms completely

---

[6]$\operatorname{diag}((\boldsymbol{R}\boldsymbol{P}_{a'|o'}\boldsymbol{P}_{oo'})(\boldsymbol{P}_{a|o} - \tilde{\boldsymbol{P}}_{a|o})^{\top}) \geq \boldsymbol{0}$ gives the matrix form of the set of $|\mathcal{O}|$ inequalities (here the function *diag* obtains the diagonal of its matrix argument).
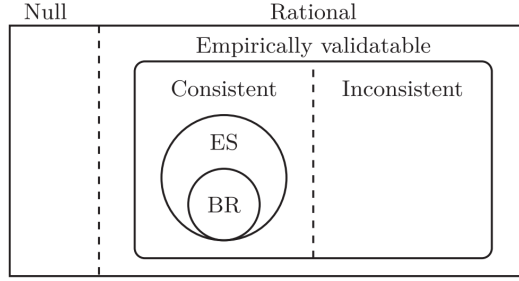
Figure 1: **Classes of norms**. Diagram taken from Morsky and Akçay [2019].

and plays the best Nash strategy). This default strategy competes with the prescriptions of norms in any selection dynamics. Norms that are externally valid are called *empirically validatable*: the prescription is rational against the prescription of another rational norm. If an empirically validatable norm is validated by itself, it is called *consistent*, else it is *inconsistent*. Within the set of consistent norms is the set of norms that are *evolutionarily stable* (ES), within which are the set of *best response* (BR) norms.

From this norm classification, we can situate correlated equilbria: a pair of norms, n and n', form the correlated equilibrium $\boldsymbol{P}_{oo'}$ and $(\boldsymbol{P}_{a|o}, \boldsymbol{P}'_{a|o})$ if n and n' are empirically validatable with respect to one another, *i.e.* eq. (5), given the matrix representation of norms. Under a natural selection process, the observed set of correlated equilibria are those arising from evolutionarily stable norms, *i.e.* norms that are stable to perturbations in norm space. Morsky and Akçay [2019] prove that such norms must be consistent, *i.e.* their prescriptions are best-responses to the opponent-environment system formed by thier descriptions, and vice versa. Next, we give an example application of this social norm theory for the game of chicken and then go on to describe an natural selection dynamics among game-playing agents by which a norm implementing a correlated equilibrium emerges.

Theory application to the canonical coordination game family: the game of chicken

The game of chicken constraints on $\boldsymbol{R}$ and the linearity of our objective (expectations of $\boldsymbol{R}$), as well as the normalization and symmetry constraints on $\boldsymbol{P}_{oo'}$ allow for a complete characterization of this canonical system.

By the linearity of the operations (taking expectations) that we will perform on $\boldsymbol{R}$, P and S can be set without loss of generality as the reference and unit of reward, respectively, so that $P = 0$ and $S = 1$. We can also parametrize temptation by the excess reward $L = T - B > 0$ of exploitation T above mutual cooperation B (the strength of *greed*; *c.f.* social dilemma condition 4b and Leibo et al. [2017]) so that the reward matrix is in general

$$\boldsymbol{R}(B, L) = \begin{array}{c} a\backslash a' \\ stop \\ go \end{array} \begin{array}{cc} stop & go \\ \begin{bmatrix} B & 1 \\ B+L & 0 \end{bmatrix} \end{array} \tag{6}$$

with condition 3 implying, $L < B - 1$, so that the exploitation excess can not overwhelm mutual cooperation[7]. We call L the *coordination challenge level* since it sets the strength of temptation relative to that of coordination.

Under the signal-following norm, $\boldsymbol{P}_{oo'}$ serves as the joint action distribution. As a result, we can parametrize it in a way that highlights performance. Its symmetry and the normalization constraint leave it with two degrees of freedom. Given the payoff matrix eq. (6), the action pair, $(stop, stop)$, gives the highest average reward (by construction, *c.f.* social dilemma condition 3). This is the case where the two players fully coordinate. We thus define the *coordination potential*, b and set $p_{\bullet\bullet} \propto b > 0$. Realizing this potential is limited by exploitation, which we parametrize using $g = p_{\bullet\bullet}$,

---

[7]The prisoner's dilemma family can be obtained from this matrix by swapping the elements in the second column and for the condition $L < B$.
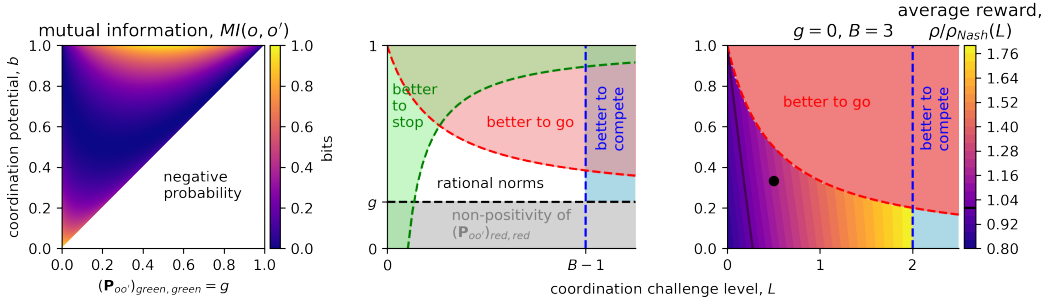
5

Figure 2: **Phase diagram for the rationality of the signal-following norm over the game family of the game of chicken**. *Left*: mutual information between environmental signals o and o' as a function the parametrization of $\boldsymbol{P}_{oo'}$, (b, g). *Center*: phase diagram for the general case. The norm is not always rational: when ● is observed, it is better to *go* when $b > (1 + 2gL)/(1 + 2L)$ (red). Similarly, when ● is observed, it better to *stop* when $b > 1 - 2g/L$ (green). Positivity of $p_{\bullet\bullet}$ requires $b > g$ (gray). When $L > B - 1$ (blue), pure competition (to exploit and be exploited) is more advantageous (*c.f.* social dilemma condition 3). *Right*: Case of $g = 0$ and $B = 3$ with average reward relative to that of the mixed Nash equilibrium (colorbar). The x-axis, L, sets the hardness of the coordination problem and the y-axis, b, sets the coordination potential. Harder coordination problems limit the coordination available. A standard version of the game of chicken ($L = 0.5$, $b = 1/3$) is shown by the black dot.

and set $p_{\bullet\bullet} \propto -g$. This gives the parametrization

$$
\boldsymbol{P}_{oo'}(b, g) = \begin{array}{c} o\backslash o' \\ \\ \bullet \\ \bullet \end{array} \begin{bmatrix} b - g & (1-b)/2 \\ (1-b)/2 & g \end{bmatrix} \tag{7}
$$

with $b > g$ by the positivity of $p_{\bullet\bullet}$. The goal then is to maximize $b - g$: getting g as small as possible, and b as large as possible. Since $b < 1$ leaks probability into the dislike pairs, coordination can be made rational even when eliminating the undesirable (go, go) action pair with $g = 0$ (we analyze this desirable case below). The pairwise correlation is $2b - 1$ and the mutual information (see fig. 2a) achieves the maximum of 1 bit at the pure symmetric $(b, g) = (1, 1/2)$ and pure antisymmetric $(b, g) = (0, 0)$ configurations.

Using the above parametrizations of $\boldsymbol{R}$ and $\boldsymbol{P}_{oo'}$, the task-environment can be completely characterized within the plane of coordination potential b and coordination challenge level L. Here, we give an example for the signal-following norm (*i.e stop* when ●; *go* when ●). We first derive the norm's rationality conditions. There is a best-response (BR) condition for each of the signal values:

received signal probability	posterior of opp.'s signal $p_{i\|j} = p(o' = i\|o = j)$	reward for playing action a when opp. follows norm ($a' \sim o'$)	rational: o is BR
$p(o = \bullet)$ $= g + (1-b)/2$	$p_{\bullet\|\bullet} = \frac{(1-b)/2}{g+(1-b)/2}$	*stop*: $\quad Bp_{\bullet\|\bullet} + 1(1 - p_{\bullet\|\bullet})$ *go*: $(B + L)p_{\bullet\|\bullet} + 0(1 - p_{\bullet\|\bullet})$	$b < 1 - \frac{2g}{L}$
$p(o = \bullet)$ $= b - g + (1-b)/2$	$p_{\bullet\|\bullet} = \frac{b-g}{b-g+(1-b)/2}$	*stop*: $\quad Bp_{\bullet\|\bullet} + 1(1 - p_{\bullet\|\bullet})$ *go*: $(B + L)p_{\bullet\|\bullet} + 0(1 - p_{\bullet\|\bullet})$	$b < \frac{1+2gL}{1+2L}$

These two rationality constraints combine with one of the positivity constraints ($b > g$) and the social dilemma condition 3 ($L < B - 1$) to define the valid region in which the signal-following norm is rational (see fig. 2b). The most important remark about this phase diagram can be made using the more simple case of $g = 0$, *i.e.* the case where the signal statistics allow for completely avoiding the punishment from mutual exploitation. We also fix $B = 3$, consistent with a widely used instance of the game of chicken (for which $L = 0.5$; see black dot in fig. 2c). The system is now parametrized by the pair (b, L). In this plane (fig. 2c), we expect that the maximum signal potential that can be realized while still ensuring it is rational to follow the signal-following norm should decrease as

6

challenge level increases, and this is indeed the case. For a given challenge level, the highest average reward is obtained at this maximum rational signal potential (where the reward for *stop* and *go* are equal when the red signal ● is observed). We have represented reward relative to that of the best (mixed) Nash strategy and see that the correlated equilibrium performs better (values greater than 1) over a large area of the parameter space (up to around 1.8 times better for L at its maximum of $B - 1 = 2$).

Note that since $|\mathcal{A}| = |\mathcal{O}| = 2$, there are 16 norms (4 possible prescriptions by 4 possible descriptions). Of the 16 norms in the game of chicken and in this case that $g = 0$ and $B = 3$, the 4 norms with prescriptions that always *stop* are null, leaving 12 rational norms (3 unique prescriptions by 4 unique descriptions). Alongside the default strategy, $\boldsymbol{P}_0$, of the mixed-strategy Nash equilibrium ($p_{stop} = 1/(1 + L)$ with average reward $(B + L)/(L + 1)$), are the 3 unique prescriptions,

$$\boldsymbol{P}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \ \boldsymbol{P}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \ \text{and } \boldsymbol{P}_3 = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}. \tag{8}$$

The identity policy, $\boldsymbol{P}_2$, is the prescription and description of the signal-following norm: just do what the signal says. $\boldsymbol{P}_1$ is the anti-signal norm: do the opposite of what the signal says. Finally, $\boldsymbol{P}_3$ prescribes that players always *go*. For prescription indexing $n = 0, 1, 2, 3$, this gives for arbitrary pair event distribution p (represented by $\boldsymbol{P}_{oo'}$) the following prescription payoff matrix

$$\boldsymbol{\Gamma}(b, L) = \begin{bmatrix} \frac{L+3}{L+1} & \frac{b-(b-1)(L(L+3)+3)+1}{2(L+1)} & \frac{-b+(b+1)(L(L+3)+3)+1}{2(L+1)} & \frac{1}{L+1} \\ \frac{L+3}{L+1} & -\frac{b}{2} - \frac{(L+3)(b-1)}{2} + \frac{1}{2} & b\left(L + \frac{3}{2}\right) + \frac{3}{2} & \frac{1}{2} - \frac{b}{2} \\ \frac{L+3}{L+1} & \frac{3}{2} - \frac{b}{2} & \frac{5b}{2} - \frac{(L+3)(b-1)}{2} + \frac{1}{2} & \frac{b}{2} + \frac{1}{2} \\ \frac{L+3}{L+1} & -\frac{(L+3)(b-1)}{2} & \frac{(L+3)(b+1)}{2} & 0 \end{bmatrix} \tag{9}$$

The highest scoring self-play norm (largest diagonal entry) is the signal following norm with reward $b(L - 2) + 3$ (the heat map in fig. 2c). This has maximums for fixed L at the boundary of the rational region where the reward for *stop* and *go* are equal when ● is received. These rewards are significantly higher than the best Nash strategy. Note that under this signal-following norm, the joint action distribution $p(a, a') = \boldsymbol{P}_2 \boldsymbol{P}_{oo'} \boldsymbol{P}_2 = \boldsymbol{P}_{oo'}$. A self-consistent interpretation of the correlating signal distribution is as the joint action distribution of another identical pair of agents or even some mean field (*i.e.* population-averaged) pair-wise behaviour. This obviates an external environment to provide the signal.

Replicator Dynamics of Social Norms

Given $\boldsymbol{\Gamma}$ as a fitness matrix, replicator dynamics is a candidate fitness-based selection dynamics among types. Applied to norms, it provides a means by which a norm can establish itself in a many-agent system [8]. It is obtained as the mean-field limit of the coupled birth-death process of N populations of player types (defined by the prescription of the norm they abide by). The system is described in this limit a continuous frequency vector $\boldsymbol{x} = (x_1, \dots, x_N)$, where $x_n \in [0, 1]$ for $n \in \{1, \dots, N\}$ is the frequency of the nth population in the system and the components satisfy the normalization constraint $\sum_n x_n = 1$. The limiting dynamics is exactly

$$\frac{\dot{x}_n}{x_n} = \Delta f_n \tag{11}$$

---

[8]We focus on rational imitation by which agents meet in pairs, compare the payoffs for each's norm, and adopt the norm with the higher payoff. A slightly more general setting to imitation that is more in the spirit of the epistemic perspective outlined in the introduction is socalled *pairwise comparison* in which the norm is adopted with some posterior probability given a noisy observation of the other player's reward. A posterior probability for unimodal noise model can be well-approximated by the logistic function,

$$p(\Delta f) = \frac{1}{1 + e^{-\beta \Delta f}} \tag{10}$$

where $\Delta f = |\Gamma_{nn'} - \Gamma_{n'n}| > 0$ is the fitness difference of the two player's norms and β is the strength of selection. This posterior results exactly from a linear log-odds model with equal prior probabilities and coefficient parameter β set as the strength of selection (e.g. noise strength in the observation model). This results in a replicator dynamics $\dot{x}_n/x_n = \tanh(\beta \Delta f_n)$ which reduces to eq. (11) in the weak selection limit (*i.e.* $\beta \Delta f_n$ typically much less than 1) for which β sets the speed of the dynamics.
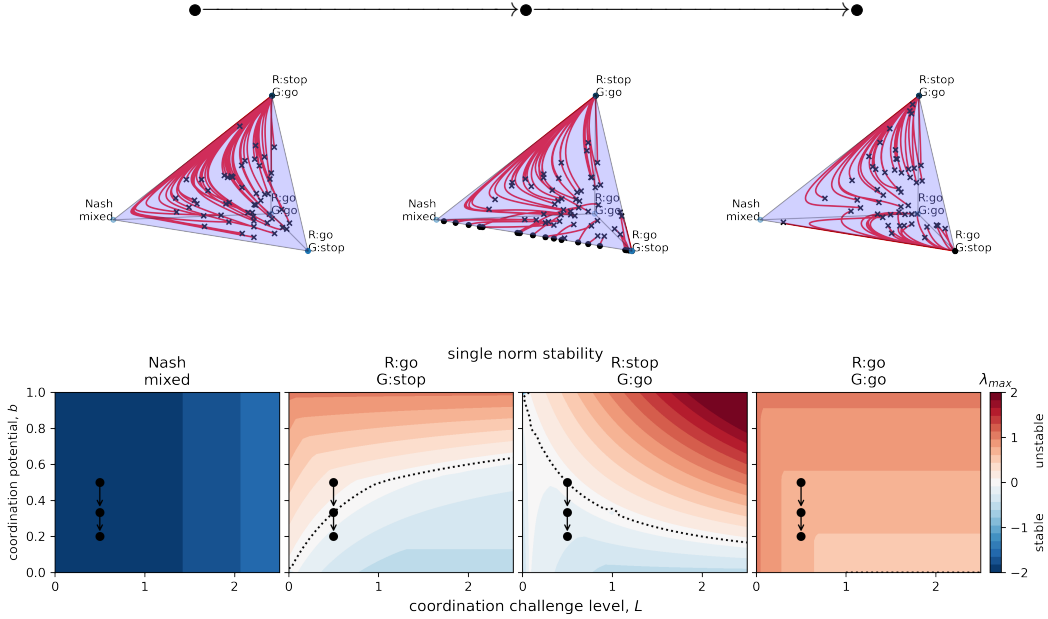
Figure 3: **Replicator dynamics (eq. (12)) for norms in the game of chicken. Top**: State space of norm frequencies in which 30 trajectories (red) are shown starting at frequencies sampled uniformly in the volume (cross markers) (left: $b = 1/2$, center: $b = 1/3$, right: $b = 1/5$). The trajectories end on the signal-following norm (black circle). Rational norms are denoted by the defining pair of $o : a$. **Bottom**: Eigenvalue spectra of the Jacobians at four single-norm fixed points over (b, L). Dotted lines denote stability transitions ($\lambda_{\max} = 0$).

with fitness deviations $\Delta f_n = f_n(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})$ from the mean fitness $\bar{f}(\boldsymbol{x}) = \sum_n x_n f_n(\boldsymbol{x})$.[9] The fitness of a type when playing against the other type (including itself) is given by the expected payoff from playing in the system $f_n(\boldsymbol{x}) = (\boldsymbol{\Gamma x})_n$, where $\boldsymbol{\Gamma}$ is the fitness matrix. Direct evaluation shows that this dynamics respects the normalization constraint. Equation (11) can be written in matrix form

$$\dot{\boldsymbol{x}} = \mathrm{diag}(\boldsymbol{x}) \left(\boldsymbol{\Gamma x} - (\boldsymbol{x}^\top \boldsymbol{\Gamma x}) \boldsymbol{1} \right), \tag{12}$$

where the function $\mathrm{diag}(\boldsymbol{x})$ evaluates to a matrix with elements of the vector $\boldsymbol{x}$ on the diagonal and 0 elsewhere and $\boldsymbol{1}$ denotes a vector of ones of the same size as $\boldsymbol{x}$. We have written mean fitness explicitly, $\bar{f}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Gamma x}$.

For the game of chicken example with 4×4 norm fitness matrix $\boldsymbol{\Gamma}$ given above eq. (9), this is a dynamics on the 3-simplex, *i.e.* the tetrahedron. Three qualitatively distinct example parametrizations of eq. (12) for the optimal signal-following norm are given in fig. 3(top).

Solution Analysis The replicator dynamics represented by eq. (12) has solutions that while complicated are tractable via standard dynamical systems analysis. Fixed points, $\boldsymbol{x}^*$ are solutions to the matrix system

$$\boldsymbol{\Gamma x} = (\boldsymbol{x}^\top \boldsymbol{\Gamma x}) \boldsymbol{1}, \tag{13}$$

where x_n is non-zero. All states at which only a single norm n is present (denoted $\hat{\boldsymbol{n}}$, *i.e.* $\hat{n}_i = \delta_{in}$) are fixed points[10]. Given the flow map $g(\boldsymbol{x})$ (the righthand side eq. (12)), stable fixed points are those fixed points $\boldsymbol{x}^*$ for which the maximum eigenvalue of the Jacobian of $g(\boldsymbol{x})$, i.e. the $N \times N$ matrix of partial derivatives denoted by $\nabla_{\boldsymbol{x}} g(\boldsymbol{x})$, evaluated at $\boldsymbol{x}^*$ is non-positive. We derive a convenient

---

[9]This dynamics is known as the generalized Lotka-Volterra equation in ecology, and is a continuous time version of the Price equation studied in the theory of biological evolution.

[10]One can also show there are no interior equilibria

matrix form here starting from the product rule,

$$(\nabla_{\boldsymbol{x}} g(\boldsymbol{x}))_{ij} = \frac{\partial g_i(\boldsymbol{x})}{\partial x_j} \tag{14}$$

$$= \delta_{ij}((\boldsymbol{\Gamma}\boldsymbol{x})_j - \boldsymbol{x}^\top \boldsymbol{\Gamma}\boldsymbol{x}) + x_i \left(\Gamma_{ij} - \left(2\Gamma_{jj}x_j + \sum_{i' \neq j}(\Gamma_{ji'} + \Gamma_{i'j})x_{i'} \right) \right) \tag{15}$$

$$= \delta_{ij}((\boldsymbol{\Gamma}\boldsymbol{x})_j - \boldsymbol{x}^\top \boldsymbol{\Gamma}\boldsymbol{x}) + x_i \left(\Gamma_{ij} - ((\boldsymbol{\Gamma}^\top + \boldsymbol{\Gamma})\boldsymbol{x})_j \right) \tag{16}$$

$$= \delta_{ij}((\boldsymbol{\Gamma}\boldsymbol{x})_j - \boldsymbol{x}^\top \boldsymbol{\Gamma}\boldsymbol{x}) + x_i\Gamma_{ij} - x_i(\boldsymbol{\Gamma}^\top \boldsymbol{x})_j - x_i(\boldsymbol{\Gamma}\boldsymbol{x})_j \tag{17}$$

$$\nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = \mathrm{diag}(\boldsymbol{\Gamma}\boldsymbol{x} - (\boldsymbol{x}^\top \boldsymbol{\Gamma}\boldsymbol{x})\mathbf{1}) + \mathrm{diag}(\boldsymbol{x})\boldsymbol{\Gamma} - \boldsymbol{x}(\boldsymbol{\Gamma}^\top \boldsymbol{x})^\top - \boldsymbol{x}(\boldsymbol{\Gamma}\boldsymbol{x})^\top \tag{18}$$

$$= \mathrm{diag}(\boldsymbol{\Gamma}\boldsymbol{x} - (\boldsymbol{x}^\top \boldsymbol{\Gamma}\boldsymbol{x})\mathbf{1}) + (\mathrm{diag}(\boldsymbol{x}) - \boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{\Gamma} - \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{\Gamma}^\top \tag{19}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Evaluating eq. (19) for single norm states $\hat{\boldsymbol{n}}$, the second term vanishes. Evaluating the first and third term, we obtain the Jacobian

$$(\nabla_{\boldsymbol{x}} g(\hat{\boldsymbol{n}}))_{ij} = (\Gamma_{in} - \Gamma_{nn})\delta_{ij} - \Gamma_{jn}\delta_{in} \tag{20}$$

$$\nabla_{\boldsymbol{x}} g(\hat{\boldsymbol{n}}) = \mathrm{diag}(\Gamma_{\cdot n} - \Gamma_{nn}\mathbf{1}) - \hat{\boldsymbol{n}}(\Gamma_{\cdot n})^\top \tag{21}$$

This sparse matrix is composed of diagonal component $\Gamma_{in} - \Gamma_{nn}$ and an nth row component $-\Gamma_{jn}$. Given $\boldsymbol{\Gamma}$, the set of eigenvalues of the single state Jacobian matrix eq. (21) can be computed numerically.

For the game of chicken example ($\boldsymbol{\Gamma}$ given by eq. (9)), the result in the (b, L) plane for the 4 single strategy states is shown in fig. 3(bottom). The always *go* norm is always unstable. The mixed Nash strategy is always stable (and relatively strongly so), but its attractor basin extends over a negligibly small fraction of the state space (frequency perturbations of magnitude 10^{-3} were sufficient to knock the system out of that equilibrium). Across a path in (b, L) on which the coordination potential is lowered, the system transitions from being the attractor basin of the signal-following norm taking up the vast majority of the state space (fig. 3top,left) to a more balanced multistability in which the attractor basin of the anti-signal following norm also takes up a significant fraction of the state space volume (fig. 3top,right). Note that at the critical value of b at which the anti-signal following norm becomes stable, a significant state space volume attracts to a population mixture of that norm and the mixed Nash strategy (fig. 3top,middle).

Discussion

We presented a detailed account of how social norms can emerge via the use of correlated equilibria.

Acknowledgments and Disclosure of Funding

References

Bryce Morsky and Erol Akçay. Evolution of social norms and correlated equilibria. *Proceedings of the National Academy of Sciences*, 116(18):8834–8839, 2019. doi: 10.1073/pnas.1817095116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1817095116.

Robert J. Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55 (1):1–18, 1987. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1911154.

Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1):40–55, 1997. ISSN 0899-8256. doi: https://doi.org/10.1006/game.1997.0595. URL https://www.sciencedirect.com/science/article/pii/S0899825697905959.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000. doi: https://doi.org/10.1111/1468-0262.00153. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00153.

Christian Hilbe. Local replicator dynamics: A simple link between deterministic and stochastic models of evolutionary game theory. *Bulletin of Mathematical Biology*, 73(9):2068–2087, Sep 2011. ISSN 1522-9602. doi: 10.1007/s11538-010-9608-2. URL https://doi.org/10.1007/s11538-010-9608-2.

Herbert Gintis. Social norms as choreography. *Politics, Philosophy & Economics*, 9(3):251–264, 2010. doi: 10.1177/1470594X09345474. URL https://doi.org/10.1177/1470594X09345474.

Robert Axelrod. An evolutionary approach to norms. *The American Political Science Review*, 80 (4):1095–1111, 1986. ISSN 00030554, 15375943. URL http://www.jstor.org/stable/1960858.

Amy Greenwald and Keith Hall. Correlated-q learning. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 242–249. AAAI Press, 2003. ISBN 1577351894.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Andrea Celli, and Tuomas Sandholm. Faster no-regret learning dynamics for extensive-form correlated and coarse correlated equilibria. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 915–916, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538288. URL https://doi.org/10.1145/3490486.3538288.

Raphael Köster, Dylan Hadfield-Menell, Richard Everett, Laura Weidinger, Gillian K. Hadfield, and Joel Z. Leibo. Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, 119(3):e2106028118, 2022. doi: 10.1073/pnas.2106028118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2106028118.

Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar A Duéñez Guzmán, Alexander S Vezhnevets, and Joel Z Leibo. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2), apr 2023. doi: 10.1177/26339137231162025. URL https://doi.org/10.1177/26339137231162025.

Michael W. Macy and Andreas Flache. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, 99(suppl_3):7229–7236, 2002. doi: 10.1073/pnas.092080099. URL https://www.pnas.org/doi/abs/10.1073/pnas.092080099.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, page 464–473, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

Petter Törnberg. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42):e2207159119, 2022. doi: 10.1073/pnas.2207159119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2207159119.

Appendix

Observing the agent population to close the loop

We detailed an evolutionary dynamics of social norms for fixed, given statistics of the signal over a defined signal space that is used to correlate behaviours in each norm. Here, motivated by the large amounts of social information in modern human experience, we instead consider the natural case that signal statistics are in fact evolving according to some dynamics derived from partial observations of the population-level behaviour itself. This self-referential choice obviates the need for an external environment to provide correlated signals by leveraging the correlations inherent in the joint action of the many-agent system.

Consider a partially observable N-agent Markov game (N even) in which agents are paired according to a (possibly stochastic) pairing $\mathcal{P}$ at each time step to play a 2-player game and where the current state is the joint action $\boldsymbol{a} = (a_1, \ldots, a_N)$ from the previous time step. One simple type of partial observation is the pair of actions from a game from the last round.[11] Whether or not pairs of agents abide by their subscribed norm however depends on whether it is rational to do so, otherwise they play (in our setting) the optimal Nash equilibrium. We assume a similar evolutionary dynamics that drives the frequencies of norm and signal pairs used in a population towards those that provide high average payoff, thus allowing the emergence of a correlated equilibria in the same way as in the previous section. The payoff matrix $\boldsymbol{\Gamma}$ now depends on these changing observation statistics

$$\Gamma_{nn'}(t) = \text{tr}(\boldsymbol{R}\boldsymbol{P}_{aa'}^{(nn')}(t)) \tag{22}$$

where

$$\boldsymbol{P}_{aa'}^{(nn')}(t) = \boldsymbol{P}^{(n')\delta}\boldsymbol{P}_{oo'}^{(nn')}(t)(\boldsymbol{P}^{(n)\delta})^{\top} \tag{23}$$

is the joint action distributions for pairings of agents with norm n and n', where $\boldsymbol{P}^{(n)\delta}$ is the conditional best-response strategy

$$\boldsymbol{P}^{(n)\delta} = \delta_{nn'}\boldsymbol{P}^{(n)} + (1 - \delta_{nn'})\text{diag}(\boldsymbol{p}_{\text{Nash}}) \tag{24}$$

where $\delta_{nn'} = 1$ if the norm $(\boldsymbol{P}^{(n)}, \boldsymbol{D}^{(n)})$ is rational given n' and $\boldsymbol{P}_{oo'}^{(nn')}(t)$ and 0 otherwise. $\boldsymbol{P}^{(n')\delta}$ is defined similarly. Finally, observations are simply empirical frequencies of actions played in the previous round of games between the two types of players, $\boldsymbol{P}_{oo'}^{(nn')}(t) = \hat{\boldsymbol{P}}_{aa'}^{(nn')}(t - \mathrm{d}t)$, where the latter is a finite-size estimator, *e.g.* for $|\mathcal{A}| = 2$, $\hat{\boldsymbol{P}}_{aa'}^{(nn')}(t) = \frac{1}{N_{nn'}}\sum_{(i,j)\in\mathcal{P}_{nn'}}\begin{bmatrix}(1-a_i)(1-a_j) & (1-a_i)a_j \\ a_i(1-a_j) & a_ia_j\end{bmatrix}$, where we use the action representation $0 \sim stop$ and $1 \sim go$ and $\mathcal{P}_{nn'} \subset \mathcal{P}$ is the subset of the pairings made for games in time step, t, from players of the two types, and $N_{nn'} = |\mathcal{P}_{nn'}|$ is the number of these pairings. When $N_{nn'}$ is large (so that finite-size fluctuations are negligible), $\hat{\boldsymbol{P}}_{aa'}^{(nn')}(t) \to \boldsymbol{P}_{aa'}^{(nn')}(t)$. By construction then, the single norm state of the signal-following norm is stable with respect to this closed-loop multi-agent dynamics.

Adding the assumption of partisan interactions leads to partisan sorting

Polarization in existing models on the effects of social media on opinion dynamics is driven by an absorbing dynamics that homogenizes opinions within groups. These models achieve differing opinions across groups with high probability by the low probability under uniform sampling of many opinions that different groups land on the same opinion. This mechanism is unrealistic in a few ways: the group opinions are uncorrelated with unrealistic parameter dependence, e.g. the fewer opinions the less polarization. A more compelling hypothesis not accounted for in these models is that the in-group homogenized opinions are *anti-correlated* across groups. Here, we propose to induce this anti-correlation as a consequence of a correlated equilibria that choreographs conflict.

Consider the following setting that implements the idea that assumed partisanship leads to partisan sorting. Each player assumes there are two agent types and know which type with which it identifies more. Players then perceive reward functions that differ for like and unlike pair interactions with

---

[11]An alternative choice is that agents observe noisy versions of the population-averaged policy (the pair actions over all games played in the previous time step).

other agents, $\boldsymbol{R} = \begin{bmatrix} B & 2\xi-1 \\ B+L & 0 \end{bmatrix}$, where $\xi = 1$ for opponents of the same type (with whom the player then plays a game of chicken) and $\xi = 0$ for opponents of the other type (with whom the player then plays a game of stag hunt so that being exploited is now more costly than mutually exploiting one another). The optimal player strategy thus depends on the opponent's type, ξ, which the player strives to infer in rational pursuit of higher reward. The basis of this inference (the computation leading to the estimate $\hat{\xi}$) lies in the inferred similarity of opinions with the opponent from partial observations of their opinions. We implement this by augmenting each player with a private belief space (represented relative to the population-averaged beliefs) as a vector $\boldsymbol{d} \in \mathbb{R}^D$ (D could be interpreted as the number of topics). Opinion similarity is defined by an opinion similarity function, $S(\boldsymbol{d}, \boldsymbol{d}')$. Here, we assume that in pursuit of inferring similarity players binarize their opinions relative to the average and use the normalized overlap $S = S(\boldsymbol{d}, \boldsymbol{d}') = \frac{1}{D}\Theta(\boldsymbol{d})^\top \Theta(\boldsymbol{d}') \in [0, 1]$ to define similarity where Θ is applied element-wise ($\Theta(x) = 1$ if $x > 0$, 0 otherwise) and set $\xi = \xi(S) = \Theta(S - 1/2)$. An opponent's opinion $\boldsymbol{d}'$ is partially observable, with single components revealed in single games. Prior assumptions about the distributions of opinions can strongly influence the inference of ξ. Under the prior assumption of non-overlapping variability of two types, $\hat{\xi} = 0$ upon the observation of any deviating opinion component. Alternatively, the maximally uninformative prior leads to a posterior with significant residual uncertainty on S from single observations that provide only a component of the opinion vector.

The different reward matrix for each pairing type drives the selection of a pair of action pair distributions via the joint-action derived signal statistics, $\boldsymbol{P}_{oo'}^+$ and $\boldsymbol{P}_{oo'}^-$. We propose that the same dynamics that leads to a correlated equilibria for the policies conditioned on playing opponents from the same population, will lead to a suboptimal equilibria when conditioned on playing against those of the different population that is characterized by choreographed conflict rather than cooperation. We also aim to show that the dependence of $\boldsymbol{R}$ on the population label can in fact be based solely on the player's estimate and thus no ground truth, type-specific populations need exist. The latter emphasizes that it is the interaction dynamics out of which the conflicted populations emerge. This is in line with recent proposals for partisan sorting via conflict alignment [Törnberg, 2022]. Under which conditions the set of local assignments are mutually consistent is then a focal question when aiming to understand the stationary behaviour of the system.