Dilemmas and trade-offs in the diffusion of conventions

Lucas Gautheron[1,2]
lucas.gautheron@gmail.com

January 28, 2025

Abstract

Outside ideal settings, conventions are shaped by heterogeneous competing processes that can challenge the emergence of universal norms. This paper identifies three trade-offs challenging the diffusion of conventions and explores each of them empirically using observational behavioral data. The first trade-off (I) concerns the imperatives of social, sequential, and contextual consistency that individuals must balance when choosing between competing conventions. The second trade-off (II) involves the balance between local and global coordination, depending on whether individuals coordinate their behavior via interactions throughout a social network or external factors transcending the network. The third trade-off (III) is the balance between decision optimality (e.g., collective satisfaction) and decision costs when collectives with conflicting preferences choose one convention. We develop a utilitarian account of conventions which we translate into a broadly applicable statistical physics framework for measuring each of these trade-offs. We then apply this framework to a sign convention in physics using textual and network data. Our analysis suggests that the purpose of conventions may exceed coordination, and that multiple infrastructures (including prior cultural traits and social networks) concurrently shape individual preferences towards conventions. Additionally, we confirm the role of seniority in resolving conflicting preferences in collaborations, resulting in suboptimal outcomes.

Keywords: conventions; collective behavior; cultural evolution; Ising model; inverse problems; simulation-based inference.

1 Introduction

Since the seminal work of David Lewis [1], conventions (including linguistic norms, technological or manufacturing standards, and many other social norms) are primarily conceived as solutions to coordination problems [2]. Yet, the present paper argues that the attitude of individuals towards conventions involves a multitude of factors beyond social coordination, resulting in tensions that may disrupt the emergence of a universal norm. To this end, we identify three trade-offs involved in the diffusion of conventions and the resolution of conflicting preferences in the absence of consensus. In addition, we show how a statistical physics approach can provide information about these trade-offs in naturally occurring scenarios. The first trade-off is the balance between i) social consistency (driven by coordination with peers), ii) sequential consistency (driven by the cost of switching from one practice to another), and iii) contextual consistency (driven by the adaptation to contextual constraints) (§1.2). The second trade-off involves the balance between *local* versus *global* coordination, depending on whether individual preferences are formed endogenously through local interactions on a network, or by factors transcending the network structure (or both, in possibly contradicting ways) (§1.3). Finally, the last trade-off is the balance between decision costs and the optimality of outcome in the

[1]Interdisciplinary Centre for Science and Technology Studies (IZWT), University of Wuppertal, Germany
[2]Département d'Études Cognitives, École Normale Supérieure, Paris, France

resolution of conflicts (§1.4). To explore these trade-offs, we develop an utilitarian account translated into a statistical physics framework which we apply to behavioral data about a sign convention in physics. This statistical framework allows us to retrieve information about the underlying coordination problem and the multiple infrastructures involved in the propagation of this convention. First, we show that scientists' attitude is driven by sequential consistency as they tend to maintain a preferred choice in their solo-authored publications independently of the target research area (§2). Then, we show that scientists' preferences are correlated – albeit imperfectly – with those of their co-authors, such that some level of coordination is achieved (§3). In order to explain how, the relative contribution of local coordination (via dyadic interactions with peers) and global coordination (i.e. via shared culture) is measured by solving an inverse Ising problem over the authors' collaboration and co-citation network. Interestingly, this approach can recover the underlying game structure while simultaneously comparing the contribution of multiple social networks to the emergence of coordination. Third, we assess the plausibility of three mechanisms of preference-formation according to their ability to explain the observed magnitudes of local and global coordination, and find slightly more evidence for a model of cultural transmission involving the imitation of peers (§4). Finally, we infer the process through which scientists resolve conflicts about which convention to use in collaborations (§5). We find evidence that the last author's preference most often prevails, thus highlighting the role of seniority and power in the resolution of coordination problems, potentially to the detriment of optimality. Taken together, these results indicate that decision-making about which convention to follow involves multiple factors that can come into conflict.

1.1 Background

While formal models of conventions provide rich insights by focusing on one or a few key features of the phenomena of interest, they may also leave out crucial aspects of reality by stripping away too much of its complexity [3], or by neglecting the interactions between phenomena studied in isolation. For instance, [4, 5] demonstrated the importance of accurately representing the topological features of complex networks (including their small-world, scale-free or clustering properties) for modeling and simulating the propagation of conventions. Similarly, while controlled experiments can uncover certain aspects of conventions in idealized settings [6–10], they may conceal the fact that complex heterogeneous processes can drive or prevent the emergence of conventions in naturalistic situations [11]. For instance, while social structures are artificially manufactured in experimental settings, one may not even know which social infrastructure is actually involved in the emergence of a convention in naturalistic situations. Fortunately, the advent of large online communities has opened up opportunities to investigate the diffusion of norms and conventions in complex networks [12], with prominent examples including Twitter [13] and Wikipedia [14]. Interestingly, to our knowledge, such data-driven approaches have not extended to the study of scientific conventions. Yet, "conventionalism" can be traced back to Poincaré, who developed geometric conventionalism as an account of the epistemic status of the axioms of geometry [15]. Additionally, conventions are ubiquitous in science, including statistical practices (such as statistical significance thresholds, which determine the level of inductive risk [16]), standard measures [17], unit systems, and technical jargon. Nevertheless, previous culturally evolutionary perspectives on science [18] have not connected formal models of conventions to empirical data from scientific settings, a gap that this paper addresses. This highlights the interactions between multiple phenomena involved in the diffusion of conventions that prior works have addressed separately or ignored, and provides cues for understanding how conventions can fail to develop into universal norms, in naturalistic settings.

Let us introduce the convention examined in the present paper. In relativistic physical theories (such as general relativity and quantum field theory), the "metric tensor" is a mathematical object that represents the metric properties of space-time. It can be seen as a matrix that defines a pseudo-distance between events according to their time and space coordinates.

In particular, in the vacuum, one can choose the metric tensor to take either of the following forms:

$$\begin{pmatrix} +1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \text{ or } \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix} \tag{1}$$

The first choice $(+,-,-,-)$ is known as the mostly minus convention (or west-coast convention) while the second choice, $(-,+,+,+)$ is referred to as the mostly plus convention (or east-coast convention). These choices are physically equivalent and lead to identical predictions. However, depending on which choice one makes, certain quantities arising in calculations will take either positive or negative values. Interestingly, there is no norm and both conventions are used. .

1.2 The trade-off between social, sequential, and contextual consistency

Below, we elaborate an utilitarian description of the effect of social, sequential, and contextual consistency on individuals' decision-making. By analogy with statistical physics, we translate this description into probabilistic models suitable for empirical exploration.

Social consistency and coordination costs Conventions are mainly conceived as solutions to coordination problems [1], which arise when individuals would benefit from acting in a mutually consistent way, but struggle to do so – maybe, for instance, because they lack the information necessary for achieving joint-action [1, 2]. Conventions can solve such coordination problems by providing individuals with expectations about how others will behave in a given setting, a paradigmatic example being left-hand versus right-hand traffic. In absence of established conventions, individuals experience *coordination costs* in their interactions. If interactions involve two people at a time, coordination costs can be represented by a payoff matrix that defines the utility (i.e. the rewards) $u_{i,j}(x_i, x_j)$ for agents i and j as a function of x_i and x_j, their respective strategies (see examples in Table 1). Additionally, coordination costs are specified by

Table 1: Examples of two-player two-action coordination games, defined by their payoff matrix. Cells indicate $(u_i(x_i, x_j), u_j(x_i, x_j))$, the rewards of i and j as a function of their joint strategy.

(a) Asymmetric coordination game ("Battle of the Sexes").

	$x_j = A$	$x_j = B$
$x_i = A$	(a, b)	$(0, 0)$
$x_i = B$	$(0, 0)$	(b, a)

(b) Symmetric coordination game under asymmetric conventions. If $a \neq b$, then the choice between A and B is less arbitrary.

	$x_j = A$	$x_j = B$
$x_i = A$	(a, a)	$(0, 0)$
$x_i = B$	$(0, 0)$	(b, b)

a network structure, which can be represented by a graph which edges' weights w_{ij} encode the frequency of interactions between any pair (i, j) of agents. More complicated games may require more complex structures (such as hypergraphs) in order to fully characterize coordination costs. In naturalistic scenarios (outside of controlled experiments), given observations of individuals' strategies, one may want to retrieve the structure of the underlying game (e.g. Table 1a versus 1b) or to infer the social network truly involved. Incidentally, it has recently been shown that coordination games (Table 1) [19, 20] can be mapped onto models from statistical mechanics, which, as we show, can be exploited for empirical explorations of conventions. To this end, one constructs a "potential" $U(x_1, \ldots, x_n)$ [21], which is a function of the joint strategy of every individual $1 \leq i \leq n$ that varies by $\sum_j w_{ij}[u_i(x'_i, x_j) - u_i(x_i, x_j)]$ as any agent i unilaterally changes their strategy from x_i to x'_i. By analogy with statistical physics, the probability of a particular combination of individual strategies can then be expressed as:

$$P(x_1, \dots, x_n) = \frac{1}{Z} e^{\beta U(x_1,\dots,x_n)} \tag{2}$$

Where Z is a normalization constant and $\beta \geq 0$ controls the degree of rationality – and efficiency – of the agents [20]. In physics, (2) is the Boltzmann distribution; U is (minus) the energy potential of a particular configuration, and β is the inverse temperature[1]. This probabilistic framework enables the retrieval of information about the coefficients of the payoff matrix $(u_{i,j})$ or the network structure (w_{ij}) from observations of individuals' strategies, as shown in §3[2].

Sequential consistency and switching costs In addition to addressing coordination problems, conventions enable individuals to settle on a specific choice among different options once and for all, in a way that facilitates future moves. Consider keyboard layouts. While there exists many such layouts (e.g. qwerty and azerty – in fact, the space of all possible keyboard layouts is very large –), we benefit from settling on one single layout, even if our choice is arbitrary and different from our peers'. In that respect, certain conventions can serve a purely internal purpose of consistency, as if individuals "played" a coordination game with themselves, such that their payoffs depend on whether their consecutive actions are mutually coherent. When switching costs are high, individuals may fail to adjust to their social environment. To model sequential consistency, let x_{it} be the convention employed by agent i at time $t \in \{1, \dots, T\}$. A simple model of the switching costs experienced by an isolated individual for a given sequence of choices is a Markov model where $U(x_{i1}, \dots, x_{iT}) = \sum_{t=1}^{T-1} u(x_{i,t}, x_{i,t+1})$, where $u(x, y)$ is the payoff associated with the transition from x to y, and $u(x, y) < 0$ for $x \neq y$. In such a model, agents experience instantaneous and non-persisting costs whenever they switch from one convention to another. Alternatively, sequential consistency may reflect lasting preferences with memory effects. In this case, instead of modeling complex long-term interactions between individual actions, one might consider the effective model $U(x_{i1}, \dots, x_{iT}) = \sum_{t=1}^{T} u_i^{x_{it}}$ where u_i^x designates the utility associated with choice x for agent i (as assumed in §2). Again, we may assume that the probability of a particular sequence takes the form $P(x_{i1}, \dots, x_{iT}) \propto e^{\beta U(x_{i1},\dots,x_{iT})}$, where β is, again, a measure of efficiency.

Contextual consistency and maladaptation costs Some conventions are less conventional than others [23, 24]. Among multiple candidates, certain *maladaptive* conventions are potentially less likely to emerge. *Maladaptation costs* can disrupt the emergence of norms if individuals are confronted to different contexts for which different conventions are best (Table 1a). Unit systems, for instance, feature context-dependent maladaptation costs. Although they do promote social and sequential coordination as other conventions, some unit systems have a small advantage in specific contexts. For instance, light-years might be a convenient unit of length for astronomers, but engineers may reasonably prefer millimeters. Maladaptation costs can be framed as a lack of consistency between a given convention and other cultural traits. This can be thought of in terms of a cultural fitness landscape [25], where $f(b_1, \dots, b_n)$ describes the fitness of a configuration of traits $\boldsymbol{b} = (b_1, \dots, b_n) \in \{\pm 1\}^n$. It is possible that the choice between, say, $b_1 = -1$ or $+1$ is "conventional", in the sense that there is no universally superior choice across the landscape (i.e. $\mathbb{E}[f|b_1 = -1] \simeq \mathbb{E}[f|b_1 = +1]$), even though certain regions in the landscape may locally favor a specific choice for b_1[3]. The utility $u_i^{x_i}$ associated with a

[1] Often, β may be omitted without loss of generality through proper rescaling of U.

[2] Individuals' behavior depends also on evolutionary rules that specify how they update their strategies. For "potential" games, the "logit" rule and the Glauber dynamics lead to the above Boltzmann distribution [21, 22]

[3] This is obvious in the context of language. The mapping between objects and symbols is highly conventional; however, for a given pre-existing language, the choice of how to name a new object can be constrained by preceding

particular choice x_i then depends on the position of i in the cultural landscape, and agents in the same region of the landscape may converge in behavior for reasons independent of social coordination (Table 1b).

Generally speaking, conventionality arises when behavior is determined by "collective" rather than individual constraints. For instance, in the case of sequential consistency, it does not matter what is the first move, as long as the *entire* sequence of actions is collectively consistent[4]. Finally, contextual consistency is also a collective constraint, since it assumes there is no way to universally reject a particular choice independently from the context in which it plays out[5].

Generally, it is plausible that all three factors are involved in conventions, albeit to varying extents. For the metric signature, coordination costs are plausible: it should be easier to collaborate with scientists who will systematically agree to using your favorite convention, and it is easier to copy results from other papers if those are systematically derived with the same convention. Switching costs are seemingly plausible, as working with different metric signatures implies keeping track of which sign certain quantities must take according to which convention is used. Finally, maladaptation costs might be involved too. For instance, for problems that involve "proper time" calculations, the mostly minus metric is advantageous, since then proper time is equal to the pseudo-distance between events rather than minus the pseudo-distance.

In §2, we start by evaluating the importance of sequential consistency and context in the case of the metric signature. It will be shown that both matter, but sequential consistency matters more, such that individuals tend to stick to their favorite convention across contexts. Therefore, physicists have *preferences* towards a metric signature. We may then ask how these preferences are formed.

1.3 Local and global processes in the diffusion of conventions

The emergence of norms is the byproduct of both "local", "dyadic" processes and pre-existing "broader population-level infrastructure" [2], including social networks or central authorities [27]. In particular, we propose to make a distinction between *local* and *global* processes of coordination. By "local" coordination, we mean, coordination emerging from local interactions on a network (e.g. by the imitation of peers [28], or strategic adjustment to their behavior), as opposed to "global" processes resulting from external factors transcending the network structure, including institutions and "central authorities" [27], cultural artifacts, as well as any pre-established cultural traits shared within different groups. Asymmetric conventions (for which certain choices are advantageous, cf. Table 1b) may propagate globally if individuals share the understanding that one option is superior. In scientific communities, local processes may propagate over a co-authorship network, while global factors may include a shared "disciplinary matrix" [29]. The local/global distinction resembles the endogeneous/exogeneous distinction made in previous works exploring the dynamics of collective attention in social media (e.g. [30]), where "endogenous" refers to behavior driven by interactions on a social network, and "exogenous" refers to processes dictated by external factors such as the mass media.

Figure 1 illustrates how local and global processes may generate different patterns of coordination. In an evolutionary game theoretic framework, locality implies that agents may only update their strategy based on the behavior of their own neighbors. In this particular example, local coordination fails to produce consensus as the network is stuck into a Nash equilibrium. Occasionally, "global" processes may solve this type of failure. Alternatively, local and global

linguistic infrastructure.

[4] One might say that the marginal probability of a particular outcome $p(x_i)$ is not constrained; only the joint probability of all outcomes $p(x_1, \dots, x_n)$ is.

[5] See epistemological holism, according to which beliefs are constrained collectively rather than in individually [26].

forces may push in opposite directions and complicate the emergence of a norm [31] – for instance, if different groups with incompatible inclinations come into contact.

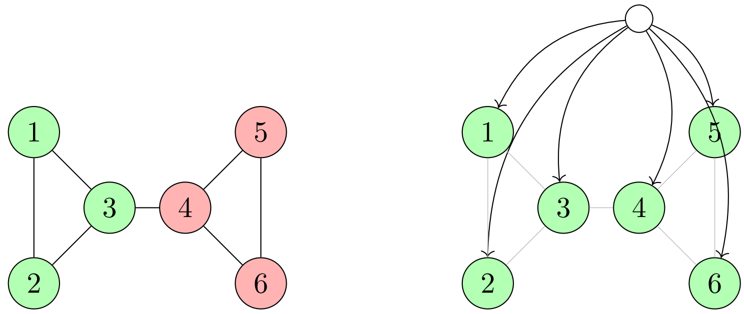


Figure 1: **Left**. Example of local coordination: nodes align to their neighbors through pairwise interactions. **Right**. Example of global coordination: nodes are coordinated by a common cause transcending the graph structure (e.g., a textbook that all authors might have learned from). Local and global processes can sometimes generate different patterns. Global mechanisms (driven by shared culture, common knowledge, or institutions) can enhance coordination in circumstances where it would be hard to achieve via local processes alone.

In §3, using an Ising model – which arises naturally from eq. (2) in coordination games [20] –, we measure the contribution of local (J) and global ($\boldsymbol{B}$) mechanisms to the formation of physicists' preferences. We find evidence for both local and global effects in the case of the metric signature, while the latter seem to pre-dominate. Moreover, as will be shown in §4, this Ising model approach allows us to compare the plausibility of more realistic mechanisms of preference-formation, according to whether they generate local or global coordination patterns. In particular, we consider three different mechanisms for the formation of scientists' preferences (without making a strong commitment to any of them), and assess their relative plausibility according to their ability to account for the measured values of J (local coordination) and $\boldsymbol{B}$ (global coordination). The first mechanism is an agent-based model that assumes physicists operate a trade-off between coordination costs, switching costs, and maladaptation costs. The second mechanism is a process of global cultural transmission (capturing, for instance, cultural transmission via textbooks) and the third mechanism considers local cultural transmission (via the imitation of peers), a channel that has the potential to propagate conventions [28].

1.4 Optimality versus decision costs in the resolution of conflicts

In the absence of a universal norm, how can coordination be achieved among individuals with conflicting preferences? Co-authorship of scientific papers provides a case in point of conflict-resolution. In order to produce a paper, authors might have to overcome disagreements about certain choices, such as which metric signature to use throughout their calculations. They must then operate a trade-off between "optimality" (e.g. the maximization of their collective satisfaction), and "decision costs" (alternatively referred to as "transaction costs" [32]). Indeed, co-authors can seek to maximize their collective satisfaction by making a collective decision, through deliberation of bargaining. However, this can be cumbersome: not all decisions deserve to be put under the whole collective's scrutiny, and it might be easier to let a leader decide, potentially at the expense of collective agreement. It is indeed well known that organizations typically develop into hierarchical structures with power and leadership asymmetries to mitigate decision and transaction costs and facilitate coordination [32, 33]. In §5, we infer the mechanisms via which physicists resolve conflicts in co-authored papers. We find some evidence that leadership also plays a role in the resolution of conflicting preferences towards the metric signature, potentially at the expense of the optimality of the decision.

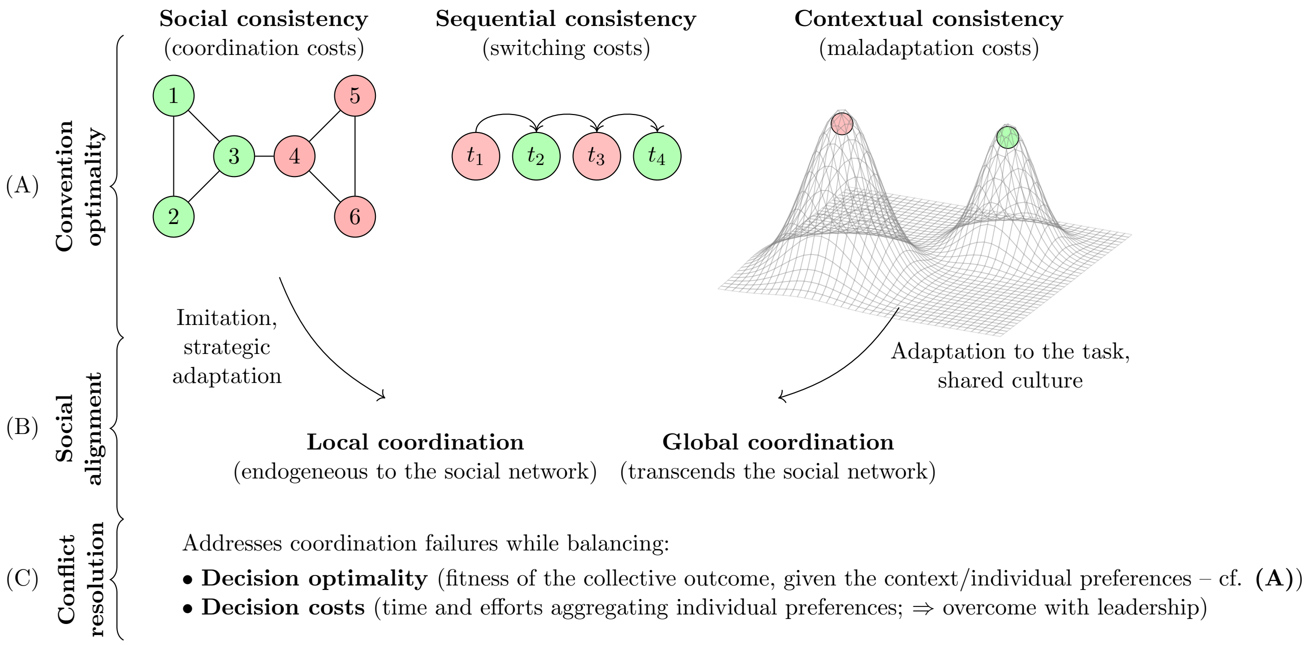


Figure 2: Three trade-offs affecting conventions and their relationships.

We have identified three trade-offs affecting conventions. Figure 2 summarises these trade-offs and highlights their interactions. In what follow, we provide empirical evidence for these trade-offs.

1.5 Data

Literature in high-energy physics is collected from the Inspire HEP database of high-energy physics, which includes various metadata (authorship, institutional affiliations, etc.). Their LaTeX source is retrieved from arXiv when available. 22500 papers from four categories (Phenomenology-HEP, Theory-HEP, General Relativity & Quantum Cosmology, and Astrophysics) are successfully classified into either metric signature (± 1) using a small set of regular expressions (see Appendix A.1).

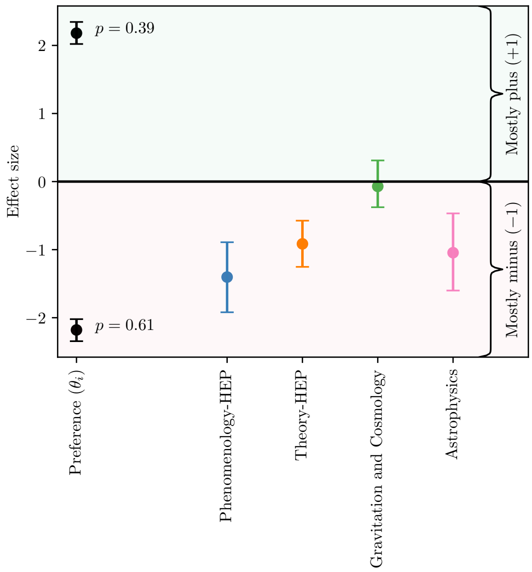
2 Beyond coordination: the role of sequential and contextual consistency

We have postulated that individuals' attitude towards conventions may be influenced not only by imperatives of coordination (i.e. social consistency), but also by imperatives of sequential and contextual consistency. If sequential consistency matters, individuals should tend to use the same convention throughout their own works. By contrast, if individuals behave differently across research areas, we may infer that they value contextual consistency.

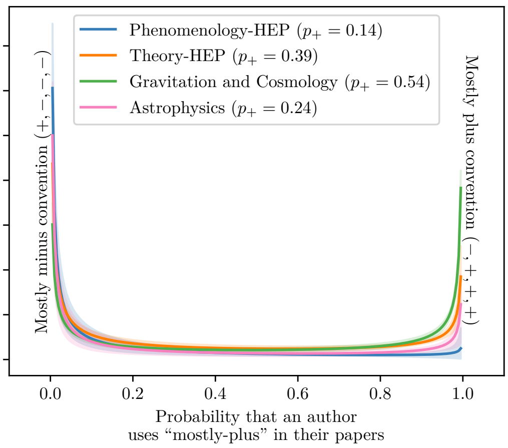
Below, we measure the importance of sequential and contextual consistency in scientists' behavior. We consider only solo-authored papers, for which the choice of metric purely reflects the sole author's choice. In order to capture the imperatives of sequential and contextual consistency, we assume that the probability that an author i uses the $+1$ sign convention in a paper d is:

$$P(\sigma_d = +1|i,c) = \text{logit}^{-1}(\theta_i + b_c) = \frac{e^{\frac{1}{2}(\theta_i+b_c)}}{e^{\frac{1}{2}(\theta_i+b_c)} + e^{-\frac{1}{2}(\theta_i+b_c)}} \tag{3}$$

where θ_i is a latent parameter that encodes author i's preference ($\theta_i > 0$ implying a preference for the $+1$ convention) and b_c is a latent parameter that encodes the bias associated with context c (the category of literature to which the paper belongs[6]). From a statistical physics perspective, this model can is the Boltzmann distribution of a physical system that can find itself in either of two states (± 1), where the degeneracy – the equivalence between these states – is "broken" by individuals' concerns for sequential and contextual consistency[7]. We assume that θ_i is drawn from a mixture of two distributions ($\theta_i = \pm\mu$), such that the model may capture the existence of two populations with a preference for each metric. We also assume that $b_c \sim \mathcal{N}(0, 1)$[8]. If $|\mu|$ is typically large, and larger than $|b|$, this would imply that scientists have preferences that generally exceed the influence of the context. As shown in Figure 3a, we find that scientists *do* have preferences that they tend to maintain across contexts, although there is some evidence that they occasionally adapt to the target research area. While we interpret such deviations from an author's preference as adaptation to the subject matter, they could indicate adaptation to the audience of the paper, in pursuit of social consistency (code-switching). Using



(a) Effect of sequential consistency (i.e. preferences, in black), and context (in color), on the choice of a convention in solo-authored papers. p indicates the prevalence of each preference (± 1).



(b) Distribution of authors' probability of using the mostly plus metric signature, according to their primary subject category. Shaded areas represent 95% credible intervals (CI). Distributions are generally bimodal, with two peaks at 0 and 1, which imply that authors tend to use always one or the other signature but rarely a mix of both in solo-authored papers.

Figure 3: **Importance of sequential and contextual consistency in scientists' behavior.**

a Beta-Binomial model[9], we confirm that authors tend to generally stick to the same metric in their works and that the prevalence of each preference varies across research areas (Figure 3b).

[6]In case a paper belongs to multiple categories, we average b_c over all these categories.

[7]This model is equivalently an item-response model [34], a popular approach in psychology and cognitive science as they allow to infer latent traits responsible for individuals' responses to a collection of tests.

[8]We assume that:
$$\theta_i = \begin{cases} +\mu & \text{with probability } p_{C_i} \\ -\mu & \text{with probability } 1 - p_{C_i} \end{cases}$$
where C_i is the primary research area of author i and $\mu \sim$ Exponential(1). The ability of this item-response model to reconstruct the latent parameters μ and b is tested with simulated data assuming no effect of sequential consistency, i.e. $\theta_i = 0$ for every author (Appendix A.2, Figure 8).

[9]Let N_i be the amount of solo-authored papers by an author i with an explicit choice of metric signature, and k_i the amount of those that uses the mostly plus ($+1$) convention. We assume that $k_i \sim \text{Binomial}(N_i, p_i)$, with $p_i \sim \text{Beta}(\alpha_{C_i}, \beta_{C_i})$ and $\alpha_c, \beta_c \sim$ Exponential(1).

3 Local versus global coordination: an Ising model approach

If scientists' attitude towards the sign convention was dictated by the need to coordinate with their collaborators, then, their preferences should be aligned with those of their social environment. While there exists no universal norm at the level of the entire field, it could still be the case that scientists are at least behaving in a way consistent with their own collaborators. To establish whether this is the case, we begin by constructing a co-authorship graph (Figure 4), where each node i on the graph (each author) is assigned an attribute $\sigma_i \in \{\pm 1\}$ that encodes their favorite convention (as measured from their solo-authored publications), and edges are given weights w_{ij} that represent the strength of the relationship between co-authors i and j[10]. We may then measure the average alignment between co-authors, $\langle \sigma_i \sigma_j \rangle = \sum_{i,j} w_{ij}\sigma_i\sigma_j / \sum_{i,j} w_{ij}$, a quantity comprised between -1 (perfect anti-alignment) and +1 (perfect alignment). We find $\langle \sigma_i \sigma_j \rangle = +0.32$, which is significantly more than would be expected by chance alone $(P < 10^{-4})$[11]: despite the absence of universal norm, scientists' preferences are positively correlated with those of their collaborators.

How did such partial alignment emerge? Coordination among physicists may be achieved either locally (via short-range interactions among scientists), or globally, via shared culture. To delineate these two possibilities, we model physicists' preferences with an Ising model, with parameters J and $\boldsymbol{B}$, such that the probability $P(\sigma_1, \dots, \sigma_n | J, \boldsymbol{B})$ of observing a particular configuration $\sigma_1, \dots, \sigma_n$ is:

$$P(\sigma_1, \dots, \sigma_n | J, \boldsymbol{B}) = \frac{1}{Z(J, \boldsymbol{B})} e^{-U(\sigma_1, \dots, \sigma_n, J, \boldsymbol{B})}, \text{ with } U = \underbrace{-\sum_{i,j} J w_{ij}\sigma_i\sigma_j}_{\substack{\text{local}\\ \text{coordination}}} - \underbrace{\sum_{i} B_{C_i}\sigma_i}_{\substack{\text{global}\\ \text{coordination}}} \tag{4}$$

Where C_i is the primary research area of i. J captures the effect of *local* coordination via pairwise interactions on the graph. $\boldsymbol{B} = (B_c)$ captures the *global* effect of each research area: their effect is global in that they equally affect all individuals within a group regardless of their position in the network. Using the cultural landscape analogy previously evoked, $\boldsymbol{B}$ can be interpreted as the "mean-field" effect of other cultural traits associated with the preference for a convention over the other, given their distribution in a particular research area. Consequently, each author experiences two influences: that of their social environment (via J), and that of their broader research area (via $\boldsymbol{B}$). In the Ising model, U is the potential of a particular configuration, as defined by the values of σ_i for every i. The Ising model follows naturally from eq. (2), §1.2 in coordination games. The $\boldsymbol{B}$ term introduces an asymmetry between authors from different research areas[12].

If $J > 0$, the potential U is lower in configurations in which each nodes share the orientation (± 1) of their neighbors. Such systems may undergo phase transitions towards configurations in which individual nodes spontaneously align over large distances. Although originated from spin physics, the Ising model provides a concise description of how local interactions at the microscopic scale can give rise to polarization at a macroscopic scale. It has found wide-ranging applications in the social science [36], notably for the study of belief propagation [37].

In our case, the configuration $(\sigma_1, \dots, \sigma_n)$ is observed, and we would like to infer the parameters of the Ising model given the data; that is, we want to extract $P(J, B | \sigma_1, \dots, \sigma_n)$.

[10] We use $w_{ij} = \sum_{d|\{i,j\}\subset A_d} \frac{1}{|A_d|-1}$, where A_d is the set of co-authors of publication d, following [35].

[11] We compare the observed value of $\langle \sigma_i \sigma_j \rangle$ to what would be expected if authors chose one or the other convention at random, with probabilities equal to the frequency of each convention. This null model predicts $\mathbb{E}[\langle \sigma_i \sigma_j \rangle] = 0.10$, far below the observed value.

[12] Unlike the games in Table 1, we assume that the effect of the asymmetry between research areas does not scale linearly with each node's degree centrality ($k_i = \sum_j w_{ij}$). Instead, each strategy is associated with a constant payoff $r_i = B_{C_i}\sigma_i$ regardless of the interactions involving i [20]

However, this probability distribution is computationally intractable. Therefore, we use the pseudo-likelihood method [38] to infer J and $\boldsymbol{B}$, given that it is both accurate, efficient, and robust to missing data as we show in Appendix A.3. The results are shown in Table 2. The inverse Ising approach reveals that research areas have large global effects, and that local coordination has a small but statistically significant effect.

In fact, this convention may propagate locally via channels others than collaborations, including publications (Figure 5). We account for this possibility by introducing an additional local contribution $J^{\text{cit}} \sum_j w_{ij}^{\text{cit}} \sigma_j$ in the approximate pseudo-likelihood (5), induced by the authors' co-citation graph G^{cit}. The co-citation graph is a directed graph that captures "who cites who" (interestingly, the pseudo-likelihood approach can directly accommodate asymmetric interactions in directed networks). Formally speaking, the weights w_{ij}^{cit} of the edges of G^{cit} measure the frequency of citations of j by i, given $w_{ij}^{\text{cit}} = \sum_{d,d'|i \in A_d, j \in A_{d'}, i \neq j} \frac{c_{dd'}}{|A_d||A_{d'}|}$ with $c_{dd'} = 1$ if d cites d' and 0 otherwise. After adding this contribution to (5), we find that both J and J^{cit} are significantly positive; that is, both co-authors and publications seem to carry an influence[13].

Table 2: Parameters of the Ising model.

Parameter	Effect size	$\text{CI}_{95\%}$	Effect size	$\text{CI}_{95\%}$
J	+0.013	[+0.009, +0.017]	+0.0095	[+0.0052, +0.014]
J^{cit}	-	-	+0.00049	[+0.00023, +0.00075]
$B(\text{hep}-\text{ph})$	-0.86	[-0.99, -0.73]	-0.77	[-0.91, -0.64]
$B(\text{hep}-\text{th})$	-0.22	[-0.29, -0.15]	-0.17	[-0.24, -0.095]
$B(\text{gr}-\text{qc})$	+0.075	[-0.0069, +0.16]	+0.076	[-0.0066, +0.16]
$B(\text{astro})$	-0.6	[-0.74, -0.47]	-0.59	[-0.73, -0.46]

To assess which of local or global coordination dominate, we evaluate the fraction of authors for which local contributions in (5) exceed the global effect of $\boldsymbol{B}$. We find that local effects exceed and reverse global effects for 7% of the sample of 2 277 authors ($\text{CI}_{95\%} = [3\%\text{–}15\%]$). In addition, we find that the inclusion of local effects only marginally improves the model's predictive accuracy, from an average of 67.7% (only considering global effects) to 70.2%. Therefore, local processes play a smaller role.

It must be stressed, however, that our measurements of J and $\boldsymbol{B}$ may be confounded by hidden structures. For instance, while we assumed $\boldsymbol{B}$ to be uniform within each of the four research areas considered, it may vary from across subtopics within each research areas. If their effect is omitted, this might inflate the estimate of J. Conversely, the effect of each research area may reflect unmodelled social structures. Therefore, the Ising model is an effective parameterization, such that the measured values of J and $\boldsymbol{B}$ may vary depending on the networks and scales under consideration. Fortunately, this approach is highly flexible. For instance, it can incorporate any combination of networks, and simultaneously infer which of these networks is more able to account for the patterns of coordination (Figure 5).

4 Inferring mechanisms of preference formation

The Ising model is certainly not a realistic description of how individual preferences come to be. Yet, as we show below, idealized models from statistical physics nevertheless provide clues about the actual process. Below, we assess the relative plausibility of three hypothetical mechanisms

[13] That J remains positive after accounting for co-citations suggests that correlations between co-authors' preferences may not be explained solely by correlations in their research.

Figure 4: **Metric signature preferences in the co-author network**. Each node is an author. Edges represent co-authorship relationships between authors. Nodes' colors indicate authors' preferences (pink for -1, green for $+1$). Only the largest connected component is shown.

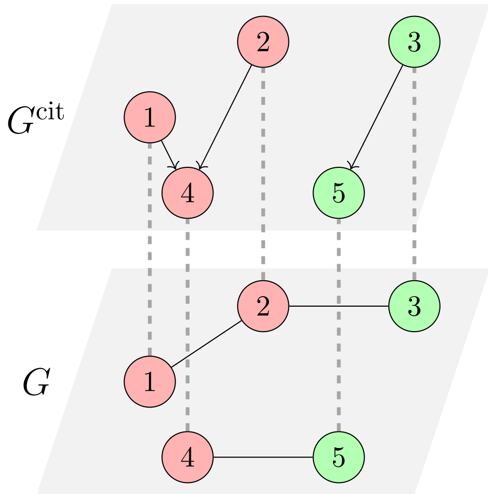


Figure 5: **Illustration of local coordination in multilayered social networks**. Nodes can be connected through different kinds of relationships (for instance, authors can be related via collaborations (G) or citations (G^{cit})). In this diagram, patterns of coordination are better explained by the directed graph at the top (G^{cit}): (1,2) have imitated (4), and (3) has imitated (5).

according to their ability account for the observed values of J and $\boldsymbol{B}$. Although none of these may be compelling accounts of reality, they will primarily serve to illustrate how to achieve some understanding of the underlying mechanism of preference-formation given the balance between local and global coordination.

The first proposed mechanism (M_1) is an agent-based model in which scientists operate a trade-off between social consistency (driven by coordination costs), sequential consistency (driven by switching costs), and contextual consistency (driven by maladaptation costs, i.e. incompatibility with their research area). In this model, the network is initialized in a random state; then, at every step of the simulation, scientists follow a best response strategy, by evaluating whether they would be better off changing their preference or not, given the magnitude of each of these costs, their probability of publishing in each research area, and their collaborators' preferences[14] (in that scenario, coordination is channeled by co-authorship and not co-citations). The second mechanism considered (M_2) is a global process of cultural transmission whereby scientists adopt a convention at the start of their career with a probability that depends on their primary research area, and on the time at which their career started. Such process is meant to capture the transmission of conventions via cultural artefacts such as textbooks (Appendix A.5). Finally, the third mechanism considered (M_3) is a process of local cultural transmission, in which scientists copy the preference of their first co-author[15].

Many samples are drawn according to each generative process M_1, M_2, M_3. For each sample, we infer the parameters of the Ising model ($\boldsymbol{B}$, J and J^{cit}) – ignoring the authors whose actual preference is unknown, in order to preserve the compatibility with the values of $\boldsymbol{B}$, J and J^{cit} inferred from the actual data). Since each model generates slightly different patterns for these parameters (Figure 6), these can be used as summary statistics for estimating their relative plausibility given the observed data, $P(M|J, J^{\text{cit}}, \boldsymbol{B})$. For this task, we use simulation-based inference [39] with BayesFlow [40, 41]. This procedure allows to perform Bayesian inference when one lacks an analytical expression for the likelihood $P(D|M)$, and all that can be done is drawing samples by simulating the generative process M. This technique is especially useful

[14]See Appendix A.4 for a more precise description.

[15]The preference of scientists with no "parent" in the graph is drawn according to the same global process as in the global cultural transmission model (M_2), such that the process M_3 includes both local and global mechanisms. In total, in this model, 10% of authors form a preference by imitation.

for making inferences about models defined by complex programs, such as agent-based models. When the data is highly dimensional (as in the present case), this approach requires "summary statistics" [39]. Interestingly, the parameters of the Ising model can serve this role. Figure 11 confirms that the procedure exhibits some ability to discriminate the three models.

The result of this procedure is shown in Figure 6. This confirms that each model predicts different patterns for J and $\boldsymbol{B}$. In particular, since it explicitly implements coordination costs (which are themselves driven by local interactions), the strategic agent model can predict large values of J. The model of cultural transmission via imitation predicts slightly higher values of J than global cultural transmission, but generally smaller values of B. Because of these distinctive patterns, we can compare each model's ability to account for the data. As shown in Figure 6, the results seem to rule out purely global cultural transmission which fails to explain the magnitude of local coordination. There is slightly more evidence of partial local cultural transmission model.

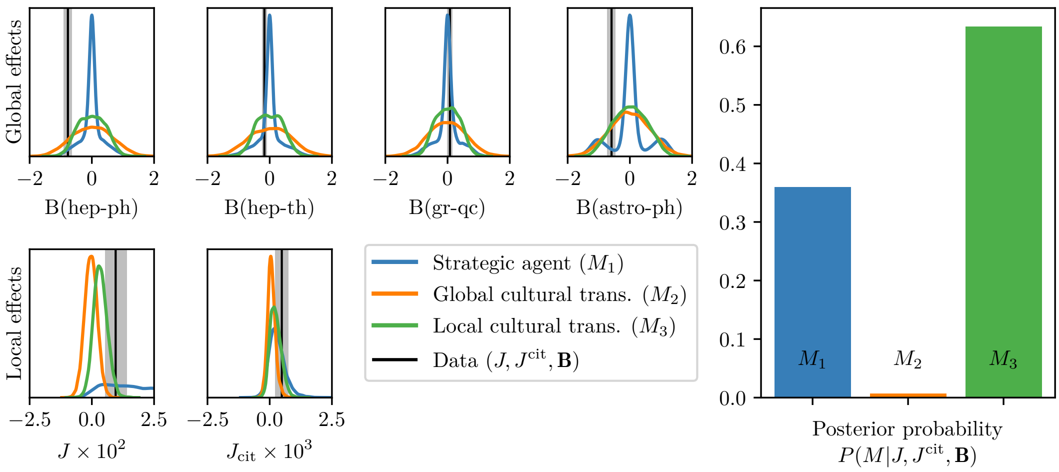


Figure 6: **Left plot:** marginal posterior distribution of summary statistics for each model (shown in colors), compared to the summary statistics derived from the data (indicated by black bars). Gray bars represent the 95% posterior credible interval of each parameter given the data. **Right plot:** posterior probability of each model given the observed parameters of the Ising model.

5 Inferring mechanisms of conflict resolution

Coordination failures give rise to conflicts. Given that physicists' preferences are not perfectly aligned to those of their collaborators, they must occasionally resolve disagreements about which metric signature to use as they co-author a paper. We stressed that the resolution of conflicts in such scenarios implied a trade-off between optimality and decision costs: while some decisions may be superior to others, the cost of arguing and properly aggregating each author's input may exceed the benefits.

Below, we consider multiple preference aggregation strategies and estimate their prevalence given data about the metric signature selected in co-authored papers. As we will show, this provides indirect information about how authors navigate this trade-off in the case of the metric signature. We leverage papers with an identified metric signature $S \in \{\pm 1\}$ for which all authors' preferences $(\sigma_i, \dots, \sigma_n) \in \{\pm 1\}^n$ were measured independently from single-authored papers. For many of these papers (182 papers with two authors, 28 papers with three authors, and 4 papers with four authors), authors have conflicting preferences. Since different processes of preference-aggregation occasionally predict different outcomes given $(\sigma_i, \dots, \sigma_n) \in \{\pm 1\}^n$, we may infer their relative likelihood from the data.

First, we consider "dictatorial" strategies, whereby a specific author (the first author, the last author, or any other one) imposes their favorite convention (which, again, is independently measured from their solo-authored publications). Dictatorial strategies dismiss all information about other authors or the research context, such that the resulting decision is potentially suboptimal. We also consider a "majoritarian" process, whereby the majority preference is selected, thus maximizing collective satisfaction. These two strategies (dictatorial and majoritarian) are probably the most classic examples in social choice theory and in the preference and judgment aggregation literature [42, 43]. It is also tempting to consider the achievement of consensus through deliberation, another popular example. However, it seems difficult to infer whether a decision was reached from deliberation based solely on the observed outcome and each individual's initial preference. Instead, we consider a "random" process, equivalent to a coin-flip (in fact, in the two-author case, a coin-flip is presumably equivalent to deliberation, if both authors are equally influential in the deliberation). Finally, we include a "conventional" process, whereby the signature most frequent in a given context is retained, irrespective of the authors' preferences.

We then estimate the prevalence prevalence π_k of each preference aggregation strategy $A_k \in \{A_1, \dots\}$, given that $P(S|\sigma_1, \dots, \sigma_n) = \sum_k P(S|\sigma_1, \dots, \sigma_n, A_k)P(A_k)$, and $A_k \sim \text{Categorical}(\pi_k)$.

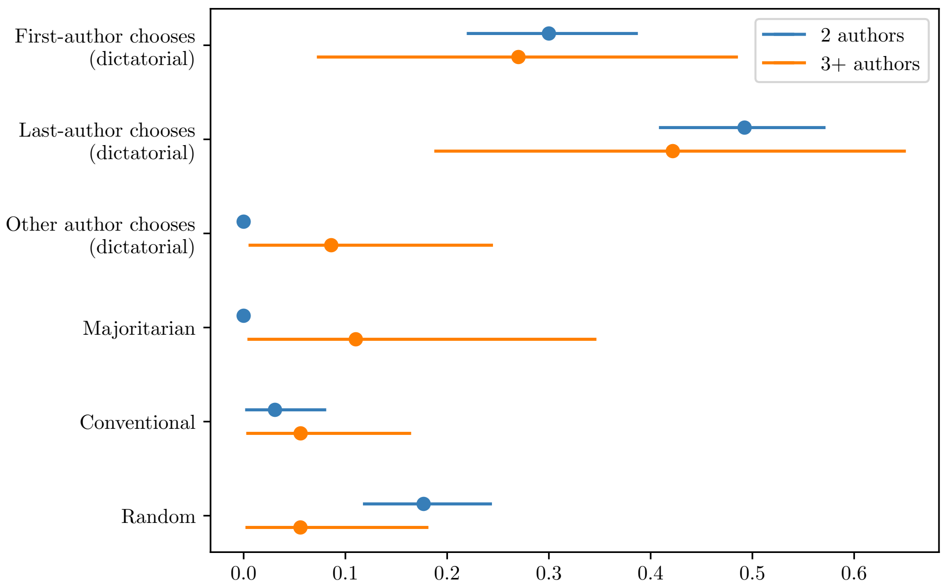


Figure 7: Prevalence of aggregation strategies. Error bars indicate 95% credible intervals. The dominant strategy seems to be that the last author dictates the metric convention.

Results are shown in Figure 7, given a flat Dirichlet prior on π_k. Due to the sample size, error bars are quite wide. Nevertheless, we can see that dictatorial strategies prevail ($\pi_{\text{dictatorial}} > 0.73$ at the 95% credible level for the two-author case and $\pi_{\text{dictatorial}} > 0.57$ for the three+-author case – which is almost always three authors), even in the 3+ authors case (for which majority vote is possible): inequalities in authors' statuses within collaborations can facilitate judgment and preference aggregation. More interestingly, in the two-author configuration, there is conclusive evidence that it is less probable for the first author to choose the metric signature compared to the last author ($P(\pi_{\text{first-author}} > \pi_{\text{last-author}}) = 0.008$). For 3+ authors, the data leans towards this direction as well ($P(\pi_{\text{first-author}} > \pi_{\text{last-author}}) = 0.222$); moreover, middle-authors seem less likely to dictate the final choice. The last author (who is generally in a leadership position) therefore seems to enjoy more influence over the choice of metric signature generally, even though the first author carries a greater share of work (in principle) and would benefit from using their favorite metric signature[16]. This emphasizes the role of leadership in the resolution of conflicts, and suggests that for this particular convention, "optimality" (whether in the sense of promoting collective agreement, or the first author's satisfaction) is sacrificed. Since the last author's preference effectively prevails more often, we may conclude that highly influential authors are protected from coordination costs and that conventions in co-authored papers are not representative of their authors' average preference.

6 Discussion

This paper identified and explored three dilemmas potentially disrupting the diffusion of conventions using a mixed theoretical and empirical approach of a sign convention in physics. This revealed that in real-life settings, the attitude of individuals towards conventions involves heterogeneous processes that may compete with each other and ultimately prevent the emergence of a norm.

The first dilemma examined in this paper is the balance between social, sequential, and contextual consistency driven respectively by coordination, switching, and maladaptation costs. In the general case, all of them may be involved and compete with each other. Conventions can thus involve more than the need to achieve coordination with others in contrast to David Lewis' account of conventions. For instance, in the case of the metric signature, we found that sequential consistency matters significantly, although physicists occasionally adapt to the topic of their research, reflecting the role of context. To investigate this trade-off, a formal and broadly applicable utilitarian description of decision-making processes involved in conventions was proposed. Building upon statistical physics, this utilitarian account was translated into probabilities, thus enabling the retrieval of information about the underlying processes from behavioral data.

In particular, for conventions ruled by coordination problems which failed to develop into universal norms, our approach can recover the underlying game (the payoff matrix) and the network structures involved. First, we confirmed that scientists' preferences tend to be aligned to those of their collaborators, although imperfectly. We then explored whether such alignment emerged from *local* coordination driven by dyadic interactions on a network, or from *global* coordination involving shared culture and knowledge or institutions transcending the social network. Interestingly, these two processes can be encoded in the structure of the underlying coordination game (Table 1). Using an Ising model, we measured their relative contribution and found significant evidence for both, although local coordination plays a smaller role. We also found that local coordination was carried by both the co-authorship and the co-citation networks. Additionally, we showed that different mechanisms of preference-formation predict different patterns for the Ising model parameters. Therefore, these parameters may be used as summary statistics to determine the relative plausibility of multiple models of preference formation. We found slightly more evidence in favor of cultural transmission of preferences via the imitation of a peer, a process that can explain a small but non-vanishing magnitude of local coordination. Purely global cultural transmission (as one might expect from the imitation of textbooks) is ruled out due to its inability to account for the observed magnitude of local coordination. However, our work did not exhaust all possible mechanisms, which was not our aim. This would require more realistic models and additional summary statistics beyond the parameters of the Ising model. For instance, different processes of preference-formation might predict different levels of intra- and inter-generational coordination, and this could be leveraged in their comparison. Nevertheless, the local versus global distinction is generally insightful. In scientific communities, it may explain which aspects of epistemic cultures belong

[16] Authorship norms are known to vary across fields [44]. To verify that these interpretation hold in fundamental physics, we evaluated the probabilities that the first-author or the last-author are strictly older than the other co-authors. We found an association between last-authorship and seniority (see Appendix A.7 for more details).

to a “disciplinary matrix” [29] (the set of practices and values that scientists adopt as part of the process of acquiring and conforming to a disciplinary identity) and which aspects emerge more spontaneously and locally. More generally, we show how the Ising approach provides a relatively model-independent way of discriminating local (i.e. emergent and endogenous) from global (exogenous) collective synchronization using behavioral network data.

Finally, given that scientists' preferences are imperfectly aligned to those of their collaborators, they must occasionally resolve conflicts about which convention to use in a collaboration. We therefore explored a trade-off in conflict-resolution between the optimality of the outcome (e.g., the degree of collective satisfaction) and decision costs (i.e. the cost of reaching a decision). We inferred the prevalence of various preference-aggregation strategies in co-authored papers, and found more evidence for “dictatorial” strategies. Specifically, we found that the last-author's preference has a higher chance of prevailing, leading to suboptimal outcomes. Therefore, leadership and seniority play a role in addressing coordination problems in the absence of norm.

By considering these three trade-offs simultaneously, we have revealed multiple interactions among themselves. For instance, whether individuals value sequential or contextual consistency has implications for the propagation of conventions and the ability to achieve consensus. Similarly, the importance of context shapes the structure of the underlying coordination game and the balance between local and global contributions in the formation of individuals' preferences. Finally, the role of seniority in the resolution of conflicts suggests that highly connected individuals can shield themselves from coordination costs, which may impact on the propagation of conventions.

Our work provides an array of tools for understanding either the lack of norm or the persistence of inferior norms and practices in a wide range of contexts. To this end, our methodology can be generalized in several ways. For instance, while the Ising model presupposes that the underlying coordination game has a dyadic interaction structure, scientists frequently interact in collaborations involving more than three authors. Therefore, we may also consider higher-order interactions [20] (encoded by hypergraphs rather than graphs) on a generalized Ising model [45]. Moreover, although this paper this paper limits itself to a binary convention, the approach can be applied to conventions involving more than two alternatives. Additionally, to our knowledge, this paper is the first attempt to reverse-engineer the processes of judgment-aggregation in co-authored publications [46], and our approach may be applied to many other decisions. Finally, this paper has not paid much attention to temporal dynamics, due to the temporal sparsity of the data. Nevertheless, exploring such dynamics would provide more information about the underlying processes of transmission, or about how sequential consistency plays out over time.

Acknowledgements Many thanks to Radin Dardashti, Thomas Heinze, Olivier Morin, and Cailin O'Connor for their feedback. I am also grateful to Jeffrey Barrett, Sean Carroll, John Miller, Scott Page, and Brian Skyrms for inspiring discussions. Finally, I would like to acknowledge the efforts of Minori Nohara, who helped with the manual annotation of a corpus of textbooks.

Code and data The code and data for this paper is available at `https://gin.g-node.org/lucasgautheron/dilemmas-conventions`.

Funding The author acknowledges funding from the DFG Research Training Group 2696 “Transformations of Science and Technology since 1800”.

Competing interests The author declares a personal inclination towards the $(+,-,-,-)$ metric signature.

References

[1] D. Lewis. *Convention: A Philosophical Study.* Cambridge, MA: Harvard University Press., 1969.

[2] R. X. Hawkins, N. D. Goodman, and R. L. Goldstone. "The emergence of social norms and conventions". In: *Trends in cognitive sciences* 23.2 (2019), pp. 158–169.

[3] C. Elsenbroich and N. Gilbert. "Modelling Norms". In: *Modelling Norms.* Springer Netherlands, 2013, pp. 143–149. DOI: 10.1007/978-94-007-7052-2_10.

[4] J. Delgado. "Emergence of social conventions in complex networks". In: *Artificial Intelligence* 141.1–2 (2002), pp. 171–185. DOI: 10.1016/s0004-3702(02)00262-x.

[5] J. M. Pujol, J. Delgado, R. Sangüesa, and A. Flache. "The role of clustering on the emergence of efficient social conventions". In: *Proceedings of the 19th international joint conference on Artificial intelligence.* 2005, pp. 965–970.

[6] F. Guala and L. Mittone. "How history and convention create norms: An experimental study". In: *Journal of Economic Psychology* 31.4 (2010), pp. 749–756. DOI: 10.1016/j.joep.2010.05.009.

[7] D. Centola and A. Baronchelli. "The spontaneous emergence of conventions: An experimental study of cultural evolution". In: *Proceedings of the National Academy of Sciences* 112.7 (2015), pp. 1989–1994. DOI: 10.1073/pnas.1418838112.

[8] R. X. D. Hawkins and R. L. Goldstone. "The Formation of Social Conventions in Real-Time Environments". In: *PLOS ONE* 11.3 (2016). Ed. by C. T. Bauch, e0151670. DOI: 10.1371/journal.pone.0151670.

[9] A. Formaux, D. Paleressompoulle, J. Fagot, and N. Claidière. "The experimental emergence of convention in a non-human primate". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1843 (2021). DOI: 10.1098/rstb.2020.0310.

[10] R. D. Hawkins, M. Franke, M. C. Frank, A. E. Goldberg, K. Smith, T. L. Griffiths, and N. D. Goodman. "From partners to populations: A hierarchical Bayesian account of coordination and convention." In: *Psychological Review* 130.4 (2023), p. 977.

[11] V. Boyce, R. D. Hawkins, N. D. Goodman, and M. C. Frank. "Interaction structure constrains the emergence of conventions in group communication". In: *Proceedings of the National Academy of Sciences* 121.28 (2024). DOI: 10.1073/pnas.2403888121.

[12] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. "No country for old members: user lifecycle and linguistic change in online communities". In: *Proceedings of the 22nd international conference on World Wide Web.* ACM, 2013. DOI: 10.1145/2488388.2488416.

[13] F. Kooti, H. Yang, M. Cha, K. Gummadi, and W. Mason. "The emergence of conventions in online social networks". In: *Proceedings of the International AAAI Conference on Web and Social Media.* Vol. 6. 1. 2012, pp. 194–201.

[14] B. Heaberlin and S. DeDeo. "The Evolution of Wikipedia's Norm Network". In: *Future Internet* 8.2 (2016), p. 14. DOI: 10.3390/fi8020014.

[15] Y. Ben-Menahem. *Conventionalism: From Poincaré to Quine.* Cambridge University Press, 2006.

[16] T. Wilholt. "Bias and values in scientific research". In: *Studies in History and Philosophy of Science Part A* 40.1 (2009), pp. 92–101.

[17] P. E. Smaldino and C. O'Connor. "Interdisciplinarity can aid the spread of better methods between scientific communities". In: *Collective Intelligence* 1.2 (2022), p. 263391372211318. DOI: 10.1177/26339137221131816.

[18] J. Wu, C. O'Connor, and P. E. Smaldino. "The Cultural Evolution of Science". In: *The Oxford Handbook of Cultural Evolution*. Oxford University Press, 2023. DOI: 10.1093/oxfordhb/9780198869252.013.78.

[19] A. Correia, L. Leestmaker, H. Stoof, and J. Broere. "Asymmetric games on networks: Towards an Ising-model representation". In: *Physica A: Statistical Mechanics and its Applications* 593 (2022), p. 126972. DOI: 10.1016/j.physa.2022.126972.

[20] F. Zimmaro, S. Galam, and M. A. Javarone. "Asymmetric games on networks: Mapping to Ising models and bounded rationality". In: *Chaos, Solitons & Fractals* 181 (2024), p. 114666. DOI: 10.1016/j.chaos.2024.114666.

[21] G. Szabó and I. Borsos. "Evolutionary potential games on lattices". In: *Physics Reports* 624 (2016), pp. 1–60. DOI: 10.1016/j.physrep.2016.02.006.

[22] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki. "Statistical physics of human cooperation". In: *Physics Reports* 687 (2017), pp. 1–51. DOI: 10.1016/j.physrep.2017.05.004.

[23] C. O'Connor. "Measuring Conventionality". In: *Australasian Journal of Philosophy* 99.3 (2020), pp. 579–596. DOI: 10.1080/00048402.2020.1781220.

[24] L. Gasparri. "Inherent and probabilistic naturalness". In: *Philosophical Studies* 181.2–3 (2023), pp. 369–385. DOI: 10.1007/s11098-023-02070-x.

[25] V. M. Poulsen and S. DeDeo. "Inferring Cultural Landscapes with the Inverse Ising Model". In: *Entropy* 25.2 (2023), p. 264. DOI: 10.3390/e25020264.

[26] W. V. O. Quine and J. S. Ullian. *The web of belief*. Vol. 2. Random House New York, 1978.

[27] H. P. Young. "The economics of convention". In: *Journal of economic perspectives* 10.2 (1996), pp. 105–122.

[28] R. Moore. "Imitation and conventional communication". In: *Biology & Philosophy* 28.3 (2012), pp. 481–500. DOI: 10.1007/s10539-012-9349-8.

[29] T. S. Kuhn. *The Structure of Scientific Revolutions*. 2nd edition, with postscript. Chicago: University of Chicago Press, 1970.

[30] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. "Dynamical classes of collective attention in twitter". In: *Proceedings of the 21st international conference on World Wide Web*. 2012, pp. 251–260.

[31] E. Lee, J. Lee, and J. Lee. "Reconsideration of the winner-take-all hypothesis: Complex networks and local bias". In: *Management science* 52.12 (2006), pp. 1838–1848.

[32] O. E. Williamson. *Markets and Hierarchies, Analysis and Antitrust Implications: A study in the economics of internal organization*. New York: Free Press, 1975.

[33] R. Calvert. "Leadership and its basis in problems of social coordination". In: *International Political Science Review* 13.1 (1992), pp. 7–24. DOI: 10.1177/019251219201300102.

[34] R. De Ayala and S. Santiago. "An introduction to mixture item response theory models". In: *Journal of school psychology* 60 (2017), pp. 25–40.

[35] M. E. Newman. "Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks". In: *Complex Networks*. Springer, 2004, pp. 337–370. DOI: 10.1007/978-3-540-44485-5_16.

[36] M. W. Macy, B. K. Szymanski, and J. A. Hołyst. "The Ising model celebrates a century of interdisciplinary contributions". In: *npj Complexity* 1.1 (2024). DOI: 10.1038/s44260-024-00012-0.

[37] M. Galesic and D. Stein. “Statistical physics models of belief dynamics: Theory and empirical tests”. In: *Physica A: Statistical Mechanics and its Applications* 519 (2019), pp. 275–294. DOI: 10.1016/j.physa.2018.12.011.

[38] H. C. Nguyen, R. Zecchina, and J. Berg. “Inverse statistical problems: from the inverse Ising problem to data science”. In: *Advances in Physics* 66.3 (2017), pp. 197–261. DOI: 10.1080/00018732.2017.1341604.

[39] K. Cranmer, J. Brehmer, and G. Louppe. “The frontier of simulation-based inference”. In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062. DOI: 10.1073/pnas.1912789117.

[40] S. T. Radev, M. D’Alessandro, U. K. Mertens, A. Voss, U. Köthe, and P.-C. Bürkner. “Amortized bayesian model comparison with evidential deep learning”. In: *IEEE Transactions on Neural Networks and Learning Systems* 34.8 (2021), pp. 4903–4917.

[41] S. T. Radev, M. Schmitt, L. Schumacher, L. Elsemüller, V. Pratz, Y. Schälte, U. Köthe, and P.-C. Bürkner. *BayesFlow: Amortized Bayesian Workflows With Neural Networks.* 2023.

[42] K. J. Arrow. *Social Choice and Individual Values.* John Wiley & Sons, 1951.

[43] C. List and P. Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford University Press, 2011. DOI: 10.1093/acprof:oso/9780199591565.001.0001.

[44] L. Waltman. “An empirical analysis of the use of alphabetical authorship in scientific publishing”. In: *Journal of Informetrics* 6.4 (2012), pp. 700–711.

[45] T. Robiglio, L. Di Gaetano, A. Altieri, G. Petri, and F. Battiston. *Higher-order Ising model on hypergraphs.* 2024. DOI: 10.48550/ARXIV.2411.19618.

[46] L. K. Bright, H. Dang, and R. Heesen. “A Role for Judgment Aggregation in Coauthoring Scientific Papers”. In: *Erkenntnis* 83.2 (2017), pp. 231–252. DOI: 10.1007/s10670-017-9887-1.

[47] M. Galesic, H. Olsson, J. Dalege, T. van der Does, and D. L. Stein. “Integrating social and cognitive aspects of belief dynamics: towards a unifying framework”. In: *Journal of The Royal Society Interface* 18.176 (2021). DOI: 10.1098/rsif.2020.0857.

A Supplementary Information

A.1 Regular expressions for determining the metric signature

The following case-insensitive regular expressions have been used to detect occurences of the mostly minus signature:

- `(([,\s\{\}]*)(\+|1)([,\s\{\}1]*)){1}(([,\s\{\}]*)\-([,\s\{\}1]*)){3,}`
- `(mostly[-\s]*minus|west[-\s]*coast)`
- `g_\{(00|tt)\}[\s]*=[\s]*[+]?[\s]*1`
- `\\Box(\^(\{2\}|2))?[\s]*\+[\s]*m\^(\{2\}|2)`

Symmetric expressions are conversely employed for detecting instances of the mostly plus metric signature.

A.2 Sequential versus contextual consistency: model assessment

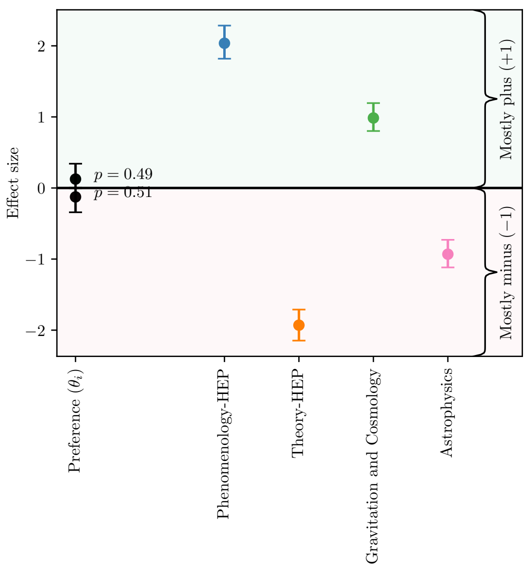


Figure 8: The analysis in 2 is re-iterated with simulated data instead of actual data. The simulation assumes that $\theta_i = 0$ for all authors (i.e. there is no effect of consistency), while each research area has a significant effect. The inference correctly finds that $|\theta|$ is nearly zero and correctly identifies the ground truth size of the effect of each research area (+2, -2, +1, and -1 respectively).

A.3 Inverse ising problem and the pseudo-likelihood approach

The pseudo-likelihood method [38] transforms the inverse Ising problem into a tractable logistic regression, based on the likelihood of observing each individual spin conditional on the others, i.e.:

$$\prod_i P(\sigma_i = +1 | \{\sigma_{j \neq i}\}) = \prod_i \frac{e^{+J \sum_j w_{ij} \sigma_j + B_{C_i}}}{e^{+J \sum_j w_{ij} \sigma_j + B_{C_i}} + e^{-J \sum_j w_{ij} \sigma_j - B_{C_i}}} \tag{5}$$

Using simulated configurations of G, we demonstrate that the pseudo-likelihood approach provides reliable estimates of J and $\boldsymbol{B}$, if all σ_j are observed, and for $J \leq 10^{-2}$ (Figure 9). In the case that a value σ_j is unknown, due to a lack of paper solo-authored by j with an identified metric signature, then author j is omitted from the sums in (5). This is equivalent to imputing $\sigma_j = 0$[17]. We find that this approach

[17] This imputation strategy is also equivalent to restricting the inference procedure to a sub-graph of the co-authorship graph, including only the nodes and edges involving the 2 277 authors whose preference could be identified in at least one solo-authored paper.

is able to recover reliable information about the true value of J (Appendix A.3, Figure 9). However, we may fear that the imputation of missing data (equivalently interpretable as the removal of unobserved nodes from the network) introduces bias in our inference [25]. A proper handling of unknown authors' preferences would require marginalizing eq. (5) over the 2^m possible combinations of the m underlying unobserved signatures[18]. Unfortunately, the amount of missing data makes this impossible. However, this issue is not necessarily critical if, ultimately, we are less interested in recovering the exact values of J and $\boldsymbol{B}$ than in using the estimates as summary statistics for the purpose of comparing multiple models of the formation of individual preferences. Then, as long as each model predicts distinct patterns for the best-fit values of J and $\boldsymbol{B}$, the procedure remains useful. In any case, simulations show that the measured value of J is very correlated with the true value, even when nodes with missing data are masked during the inference process (cf. Appendix A.3, Figure 9). Finally, missing data could be a feature rather than a bug; they might manifest the fact that certain authors make no explicit use of a specific metric signature, in which case it is reasonable to assume that they may not exert any influence over their co-authors' preferences.

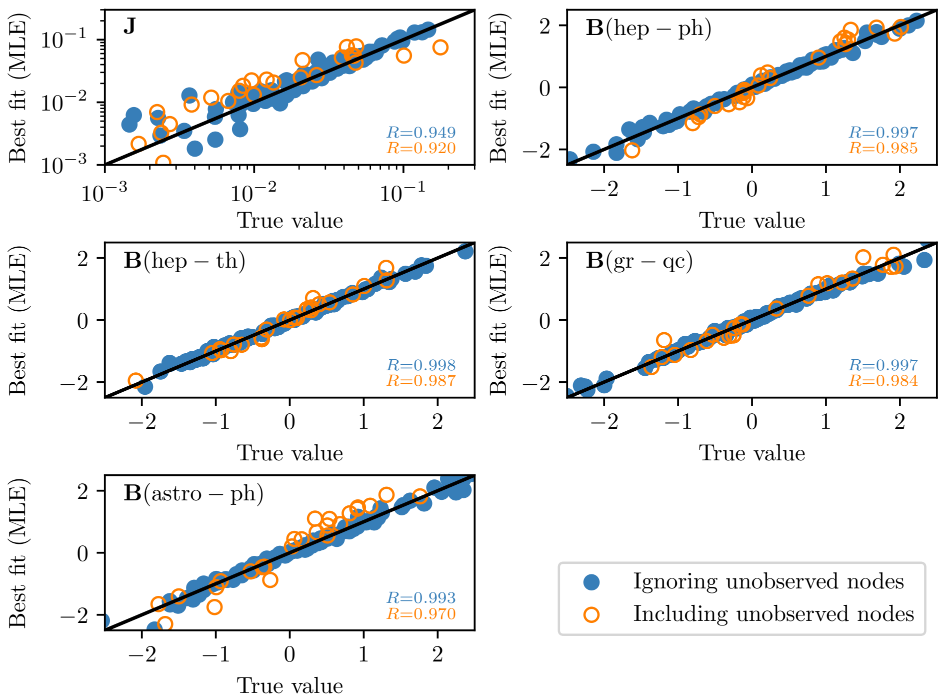


Figure 9: Robustness of the pseudo-likelihood approach for measuring J and $\boldsymbol{B}$. "True" values of J and $\boldsymbol{B}$ are drawn at random [$J \sim \text{Exponential}(1/J^*)$, $\boldsymbol{B} \sim \mathcal{N}(0,1)$]. Node configurations (σ_i) are drawn at random according to the Ising model for each values of J and $\boldsymbol{B}$, using Gibbs sampling, either i) removing or ii) including nodes corresponding to authors whose preference is not observed in the data. Finally, the maximum likelihood estimates (MLE) J^{MLE} and $\boldsymbol{B}^{\text{MLE}}$ are recovered with the pseudo-likelihood approach, for each configuration (σ_i), imputing $\sigma_i = 0$ for authors whose preference was not observed in our data. The best-fit values are in reasonably good agreement with the true values over the simulated range, although they are much less accurate in the case where unobserved authors are included in the Gibbs sampling process.

A.4 Strategic agent model

The "strategic agent" model proceeds as follow:

1. The parameters of the model are drawn at random:

[18] An alternative would be Gibbs sampling, which may handle missing data without marginalization, though it turned out to perform worse than HMC in the present case.

- $c_b \sim \mathcal{N}(0,1)$, defined for each research area b, is the (dis)advantage of the $+1$ convention in b. The cost of using a convention σ in context b is $\max(0, -\sigma c_b)$.
- $c_c \sim \text{Exponential}(\langle d_i \rangle)$ represent the magnitude of coordination costs, where $\langle d_i \rangle$ is the average degree-centrality of authors in the co-authorship graph. The mean is thus set such that $\langle c_c \rangle \langle d_i \rangle = 1$.
- The cost of switching from one convention to another is fixed ($c_s = 1$)[19].

2. At $t = 0$, the network is initialized in a random state: $\sigma_{i,t=0}$ is set to either -1 or $+1$ with equal probabilities.
3. At $t+1$, each agent compares their payoff in two scenarios: i) they switch their preference ($\sigma_{i,t+1} = -\sigma_i$) or ii) they maintain it ($\sigma_{i,t+1} = \sigma_i$). The difference in payoffs is:

$$\Delta = -c_s - c_c \sum_j w_{ij} \left(\max(0, \sigma_{j,t}\sigma_{i,t}) - \max(0, -\sigma_{j,t}\sigma_{i,t})\right) - \sum_b p_{ib} \left(\max(0, \sigma_{i,t} c_b) - \max(0, -\sigma_{i,t+1} c_b)\right) \tag{6}$$

   Where p_{ib} is the probability that i publishes in research area b. If $\Delta > 0$, i switches their preference. The cost of switching (c_s) introduces an asymmetry in Δ and has the effect of a conservative bias.
4. The process is repeated 50 times. The amount of steps reflects a compromise between performance and convergence.

This best-response strategy model is similar to common logit-response approaches to belief dynamics such as [47], in the limit $\beta \to +\infty$ (see eq. 1.6).

A.5 Global transmission model

For the global transmission model, we assumed that the probability of adopting a specific convention depends on both time and the author's primary research area. The time-dependence was captured by a random walk. The rate of change in the random walk was obtained by fitting the model to data on reference books for which approximate patterns of citations throughout time could be measured. We manually determined the metric convention used in each of these references. These gave us a measure of the prevalence of each convention in the citations of reference textbooks' throughout time. Unfortunately, this measure itself was too imperfect to reflect the actual probability that a scientist adopts a convention from a specific textbooks. Nevertheless, we used the rate of variation of this measure with time in our random walk model.

A.6 Distribution of summary statistics across models

Conditioning the outcome of simulations on high-dimensional data D to evaluate $P(\cdot|D)$ is difficult because the probability of generating exactly D becomes virtually zero. One should therefore condition on summary statistics T living in a lower dimensional space. Ideally, the mapping $f : D \mapsto T$ should be chosen in a way that maximizes our ability to tell apart the hypotheses that we seek to discriminate. In our case, $f : (\sigma_1, \dots, \sigma_n) \mapsto J, \boldsymbol{B}$ may not be optimal in that specific sense, but it has *some* discriminating power (see Figure 11) and has the merit of interpretability. A trivially better summary statistic for assessing the plausibility of, say, the model of local cultural transmission would be, for instance, the average rate of agreement between each author's and their first co-author (whose preference they should have imitated, according to the model).

[19] This breaks a degeneracy of the model due to scale-invariance (if all costs were rescaled by a certain quantity, agents' behavior would remain identical).

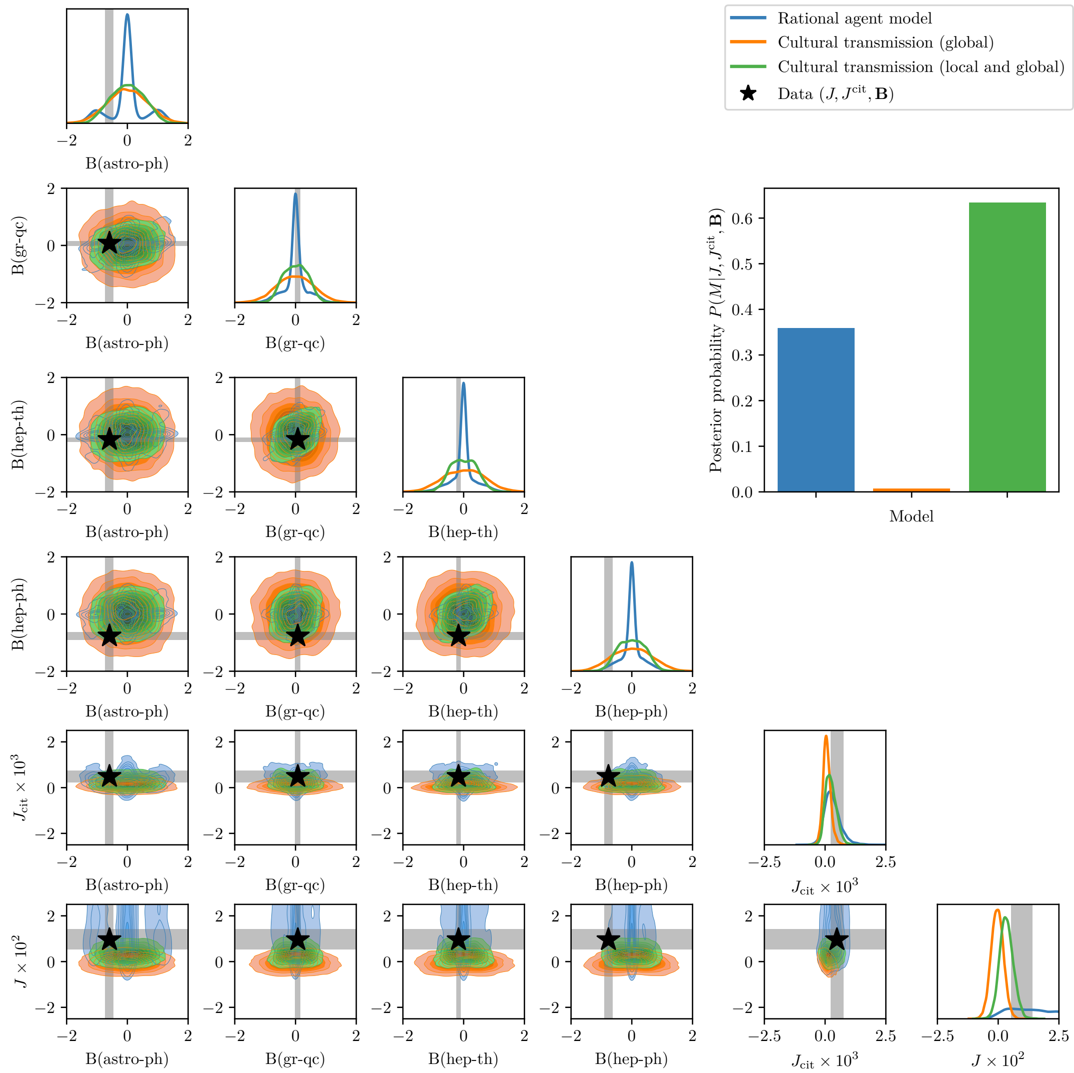


Figure 10: **Bottom-left pair plot:** distribution of summary statistics for each model (shown in colors), compared to the summary statistics derived from the data (shown as black stars). Plots on the diagonal show the marginal posterior distribution of each summary statistics for each model (gray bars represent the 95% posterior credible interval of each parameter given the data). **Top-right bar plot:** posterior probability of each model given the observed parameters of the Ising model.

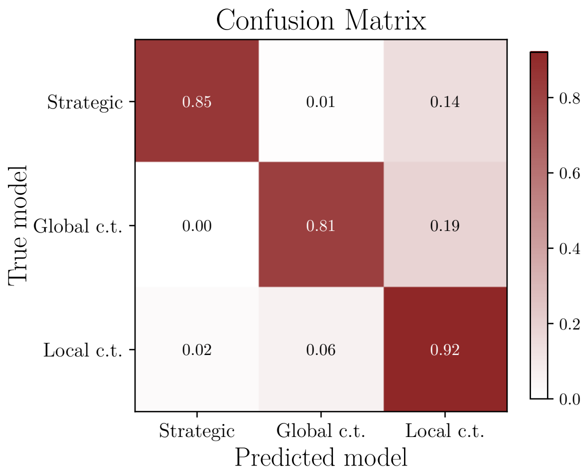


Figure 11: Reliability of the simulation-based model comparison approach. The confusion matrix represents the probability that a sample drawn from the horizontal model is attributed to the vertical model.

A.7 Authorship norms

We investigated authorship norms in fundamental physics (excluding experimental physics, which are not considered in this paper and have very unusual norms). We found that the author-list of 79% of two-author papers are alphabetically ordered. Given that for n authors, there is a $1/(n!)$ chance that any ordering is equal to the alphabetical order, this implies that 56% of two-author papers author-lists are *intentionally* ordered [44]. This number goes down to 45% for four-author publications. Therefore, despite a high prevalence of alphabetical ordering in fundamental physics compared to other disciplines (as found by [44]), in about half of the publications the ordering of authors is meaningful.

Most importantly, we found evidence that last-authorship is associated with seniority: in 54% of two-author papers, the last author has an academic age strictly higher than the first author; in comparison, in only 40% of cases, the first-author has strictly higher seniority compared to the last-author. In the three-author case, the last author has the strictly highest seniority in 29% of cases, versus 17% for the first-author.