Noname manuscript No.
(will be inserted by the editor)

Establishing Norms with Metanorms in Distributed Computational Systems

Samhar Mahmoud · Nathan Griffiths · Jeroen Keppens · Adel Taweel · Trevor J. M. Bench-Capon · Michael Luck

the date of receipt and acceptance should be inserted later

Abstract Norms provide a valuable mechanism for establishing coherent cooperative behaviour in decentralised systems in which there is no central authority. One of the most influential formulations of norm emergence was proposed by Axelrod [2]. This paper provides an empirical analysis of aspects of Axelrod's approach, by exploring some of the key assumptions made in previous evaluations of the model. We explore the dynamics of norm emergence and the occurrence of norm collapse when applying the model over extended durations. It is this phenomenon of norm collapse that can motivate the emergence of a central authority to enforce laws and so preserve the norms, rather than relying on individuals to punish defection. Our findings identify characteristics that significantly influence norm establishment using Axelrod's formulation, but are likely to be of importance for norm establishment more generally. Moreover, Axelrod's model suffers from significant limitations in assuming that private strategies of individuals are available to others, and that agents are omniscient in being aware of all norm violations and punishments. Because this is an unreasonable expectation, the approach does not lend itself to modelling real-world systems such as online networks or electronic markets. In response, the paper proposes alternatives to Axelrod's model, by replacing the evolutionary approach, enabling agents to learn, and by restricting the metapunishment of agents to cases where the original defection is observed, in order to be able to apply the model to real-world domains. This work

Samhar Mahmoud
King's College London E-mail: samhar.mahmoud.ac.uk

Nathan Griffiths
University of Warwick E-mail: nathan.griffiths@warwick.ac.uk

Jeroen Keppens
King's College London E-mail: jeroen.keppens@kcl.ac.uk

Adel Taweel
King's College London E-mail: adel.taweel@kcl.ac.uk

Trevor J. M. Bench-Capon
University of Liverpool E-mail: tbc@liverpool.ac.uk

Michael Luck
King's College London E-mail: michael.luck@kcl.ac.uk

can also help explain the formation of a "social contract" to legitimate enforcement by a central authority.[1]

1 Introduction

Modern state-based societies are reliant on a complex system of abstract legal norms that govern the behaviour of individuals and a central authority that enforces these norms. In his seminal work on the sociology of law, Max Weber has argued that these are prerequisites for such societies to exist [38]. Indeed, legal norms enforced by a central authority enable individuals to specialise their roles in society, which, in turn, enables the division of labour that industrialisation and technological advancement of our societies require [18]. The field of AI and Law is largely concerned with intelligent systems that reason with or about such norms and support the work of agents of the central authority that enforces them.

Aside from centrally enforced abstract legal norms, Ehrlich has identified a system of so-called "living law", an assortment of customs and moral norms enforced through social relationships between individuals [19]. Although living law does not involve the complex enforcement infrastructure of legal norms, it too governs the behaviour of individuals in society. Such living law may be the precursor to legal norms. Occasionally, living law has been a precursor to centrally enforced legal norms. Druzin, for example, argues that much of commercial law has evolved from the customs of merchants and was codified into a formal legal system by states at a later stage [17]. Interestingly, Druzin suggests that frequent interactions, as is the case in commerce, is conducive to the emergence of norms that are incorporated in legislation eventually, and his work examines the effect of engagement on norm emergence.

In distributed computation systems in general, and internet systems in particular, there exist an increasing number of communities consisting of interacting individual human and/or machine agents, typically with relatively little control from a central authority. Often, the volume, frequency and nature of interactions inhibit effective functioning of a central authority (or, as in pioneer societies, the central authority is too remote for it to be feasible for it to enforce the norms) and, sometimes, a central authority is not deemed desirable for the community. Nonetheless, these communities too stand to benefit from a system of norms that governs interactions between individuals, though the norms need to emerge spontaneously from the community. This paper presents an empirical study of the conditions that are conducive for such a "legal order" to emerge in such distributed computational systems.

In many application domains, engineers of distributed systems may choose, or be required, to adopt an architecture in which there is no central authority, and the overall system consists solely of self-interested autonomous agents. The rationale for doing so can range from efficiency reasons to privacy requirements. In order for such systems to achieve their objectives, it may nevertheless be necessary for the behaviour of the constituent agents to be cooperative. In peer-to-peer file sharing networks, for example, it is required that (at least a proportion of) peers provide

[1] Michael Luck gave a keynote talk at the Fifteenth International Conference on AI and Law, part of which was based on the work reported in this paper.

files in response to the requests of others, while in wireless sensor networks nodes must share information with others for the system to determine global properties of the environment. However, there is typically a temptation in such settings for individuals to deviate from the desired behaviour, which is known as the free rider problem [24,29,59]. For example, to save bandwidth, peers may choose not to provide files, and to conserve energy, the nodes in a sensor network may choose not to share information. Therefore, some form of mechanism is needed to outweigh such temptations and to encourage cooperation among self interested agents.

As has been suggested by many (e.g., [7,8,11,16,23,33,36]), *norms* provide a valuable mechanism for regulating or constraining human societies. Perhaps the most obvious and clear manifestation of norms is when they arise through the explicit introduction of laws that are established by legislatures, for example, or through rules or bye-laws of smaller groups such as member clubs. However, norms are also valuable when there is no central authority, and they emerge as a result of individual behaviour, in order to establish some coherence or stability in a group. It is this latter aspect that has been the focus of several researchers, such as [26, 30,35,46,52].

In studying emergence, for example, Epstein [20] proposes a model in which agents decide which side of a road to drive on in a particular area by imitating the decision made by the majority of other agents in the area, while Borenstein and Ruppin [9] study the use of learning by imitation to enhance the evolutionary process through lifetime adaptation. Similarly, Savarimuthu et al. [43] use imitation in considering the *ultimatum game* in the context of providing advice to agents on whether to change their norms in order to enhance performance. In other work, Shoham and Tennenholtz [48] propose the use of the highest cumulative reward (HCR) learning algorithm, by which agents learn to choose the action that brings the highest reward. Kittock [28] studies the iterated cooperation game (ICG) [27] and the iterated prisoner's dilemma (IPD) [1], both of which are iterative games in which two agents choose one of two actions, with their reward depending on the combination of their choices. Urbano et al. [53] use a classical convention emergence game to study the influence of interaction topologies (random, regular, small world and scale-free topologies). More recently, Franks et al. [21] have shown that inserting a small amount of influencer agents — those with specific conventions and strategies — is enough to manipulate the convention adopted by large societies.

The models used in these previous efforts are relatively unsophisticated, with only two agents involved in a single interaction, making them difficult to use in regulating agent behaviours in domains of our interest, such as peer-to-peer file sharing and wireless sensor networks, in which there are many interacting agents. However, Axelrod's seminal paper in 1986 offered a model of norms and metanorms [2] (that has since been investigated further [22,37]), in which multiple agents interact in a way that suggests potential application to such domains. Axelrod's model is a game in which different agents decide whether to defect or cooperate (comply). Agents may also observe others and have the ability to punish those who defect. Here, an agent's behaviour is assessed by means of a careful scoring system that simulates the potential rewards and penalties associated with norm violation and enforcement. Key to Axelrod's model is the notion of *metanorms*, secondary norms that help to enforce compliance with primary norms by punishing agents that fail to punish a defector. By using metanorms, Axelrod was able to establish norms in his experiments. Note that by introducing metanorms, the situation is one in which punishment of defection becomes a duty, rather than one in which it is simply permitted. By accepting metanorms, the agents can see themselves as part of a wider society, with obligations to that society as a whole, which paves the way for central enforcement to be accepted, and even welcomed.

Axelrod's model may be a solution to our problem but, as was more recently shown by Galan and Izquierdo [22], his results are dependent on both certain assumptions and some very specific and arbitrary conditions. In this paper, we elaborate on the work of Galan and Izquierdo by analysing the results of Axelrod's model, showing that Galan and Izquierdo's results also rely on some particular assumptions and conditions. We also provide a further analysis of Axelrod's model, drawing out some important considerations for the establishment of norms more generally. In addition, various improvements are needed in order to generalise the metanorm approach, and make it applicable over *distributed* computational systems. The model assumes a central authority and full access to agents' private strategies, which is difficult to justify in real world distributed computational systems. In response, this paper investigates alternatives that allow us to make use of the mechanisms resulting from Axelrod's investigations, in more realistic settings.

The remainder of this paper is organised as follows. Section 2 introduces related work. Section 3 provides a description of Axelrod's model followed, in Section 4, by a more detailed analysis of the results than provided elsewhere, leading to a new consideration of the circumstances for norm collapse (when norms are not established). Section 5 introduces limitations of Axelrod's model that are revealed from the analysis of the results. In Section 5.1, we present a new strategy copying technique, and show how it performs in the original context and in situations in which observation of defection is not guaranteed. In Section 5.2, we describe a reinforcement learning algorithm designed to avoid the need for access to the private strategies of others. Finally, we present our conclusions in Section 6.

2 Related Work

The existence of norms does not mean that agents will always comply with them. Within the literature, norm compliance can be established using two different approaches, top-down and bottom-up. In the top-down approach, a norm is imposed by some kind of authority, which is responsible for monitoring compliance with the norm. The impact of such centrally imposed sanctions, promulgated by an authority, in a norm regulated multi-agent setting has been discussed in some detail in [3]. In the bottom-up approach, in contrast, agents become aware of the existence of a norm through their interactions with each other, and it is the responsibility of every agent to monitor compliance with this norm. The top-down approach is sometimes referred to as *normative reasoning*, while the bottom-up approach is often referred to as *norm emergence* and is the focus of the work presented in this paper. While there has been a considerable amount of work in the area of norm emergence [26,30,35,46,52], to focus the paper appropriately, we introduce only the most closely related research on using imitation or learning to study emergence with the presence and absence of sanctions. For a more general overview, the work of Savarimuthu et al. [44] provides an excellent reference.

Normative games have been widely used as environments in which various mechanisms that support norm emergence can be evaluated (for example, [52]).

Conventions and norms are related concepts, and the terms are sometimes (especially in a multi-agent systems context) used interchangeably. Of course, it is possible to draw distinctions, and there is a considerable literature on the distinction in the humanities and social sciences (for example, [4]).

Often, norms involve the possibility for some kind of sanction or punishment for violation, or reward for adherence, while conventions are seen as behaviours or strategies that are adopted without these additional mechanisms, but this is by no means a clear cut difference, since some norms do not have sanctions attached (for example, some moral norms), whereas some conventions are reinforced with sanctions, such as mockery or social ostracism. For our specific purposes in this paper, the distinctions are not crucial. What we require is that the agents have a code of behaviour (which could comprise norms, conventions or even laws), which they may conform to (*cooperation* in game theory terms) or deviate from (*defection* in game theory terms). Deviation has a positive payoff for the agent deviating and a negative payoff for other agents. Deviation makes an agent liable to punishment, which has a cost both for that agent and the agent punishing. When we introduce *metanorms*, failure to punish becomes a deviation itself liable to punishment. We give more precision to these notions in the next section when we discuss Axelrod's normative game [2]. In this paper, we will mostly use the term *norms* for the constituents of this code of behaviour, since that is the term most often used in agent systems, which is where our research background lies. No particular weight should be placed on the term, however, and when we discuss work that has used a different term, we use their term. Most of the systems we discuss come from the field of artificial intelligence (AI) and multi-agent systems, although there has been some related work specifically in AI and law (for example [40, 41]).

Some researchers [49, 51, 55] have focussed on *convention emergence*, a process that starts with a population of agents with no initial preferences for any of the existing conventions. These agents must agree on a common convention (which is, by definition, only useful when all agents share the same one); once all agents have agreed on the same convention, the system has converged, and we can say that the convention has emerged. A classic example of convention emergence games is the iterated cooperation game (ICG) [27]. Salazar et al. [42] propose a variation of the coordination game (as in *convention emergence*), implementing a *language coordination game* similar to Lakkaraju and Gasser [31]. In this game, agents must match words with concepts, creating a huge convention space.

The *ultimatum game* [50] is a game in which two agents interact with each other in order to share money. The first agent proposes to the second how to share the money, and the second is free to accept or reject the proposal. If accepted, each agent takes its share, otherwise neither gets anything. Here, each agent has a personal norm that defines its proposal strategy. Villatoro et al. [54] study the *emergence of cooperation*, where there is a conflict between the individual and the collective interests, making this the main difference with respect to *conventions*. Here, actions are not chosen randomly, because the targets' interests lie in the direction of action opposing compliance with the norm, and the beneficiaries' interests lie in the direction of action favouring compliance with the norm. The iterated prisoner's dilemma (IPD) [1] can be seen as a classic example of this. In this context, Lloyd-Kelly et al. [32] investigate the effect of simulated emotions on the emergence of cooperation among a population of agents playing IPD. Their model incorporates various emotional attributes, such as anger and gratitude, which influence agent decisions to cooperate or to defect and they also consider the possibility for social interventions to prevent agents from continuous defections. Simulation results suggest the best intervention strategy to use for each such emotional attribute.

Within the context of the previous normative games, various forms of learning have been an effective means for facilitating norm establishment among groups of self-interested agents. Learning by imitation has been used by various researchers [25, 13] in artificial intelligence, and by Epstein [20] in the context of norm emergence. In Epstein's model, agents must decide which side of a road to drive on, where the decision of each agent is determined by observation of which side of the road already has more agents driving on it, within a particular area. In this respect, agents imitate what the majority of their neighbours are doing. Borenstein and Ruppin [9] study the use of learning by imitation to enhance the evolutionary process through lifetime adaptation and show that imitation successfully achieves such a goal.

Similarly, Savarimuthu et al. [43] use imitation in considering the *ultimatum game* in the context of providing advice to agents on whether to change their norms in order to enhance performance. Here, each agent has a personal norm that defines its proposal strategy. In addition, agents are able to request advice regarding their proposal strategy from only one agent, the *leader*, which is believed to have the best performance in the requesting agent's neighbourhood. Moreover, agents are capable of accepting or refusing the advice according to their autonomy level.

Walker and Wooldridge [57] use a simple strategic update function in their model, based on the work of Conte and Castelfranchi [12]. In their model, agents wander around searching for food in order to gain energy. However, since this movement causes them to lose energy, they need to find as much food as they can, and expend the least energy in doing so. For this reason, agents follow different strategies, and change from one strategy to another according to a *majority rule*, which instructs an agent to switch to another strategy if it finds that the other strategy is used by more agents than its current strategy. Shoham and Tennenholtz [48] propose the use of the highest cumulative reward (HCR) learning algorithm, by which agents learn to choose the action that brings the highest reward. HCR has also been used by Delgado et al. [15] in the context of a simple framework in which an agent makes a choice between two different actions and receives a positive payoff if they both choose the same action, or a negative payoff if their actions are different. Agents record the outcome of taking each of the two actions and pick the action with the better outcome for the next interaction.

A more complex form of learning has been used by Mukherjee et al. [34] and Sen and Airiau [46], who adopt Q-learning and some of its variants (WOLF-PHC and *fictitious play*) to show the effect of learning on norm emergence. They experimented with two different scenarios, first with homogeneous learning agents (where all agents have the same learning algorithm), and second with heterogeneous learning agents (where agents can have different learning algorithms). Their results suggest that norm emergence is achieved in both situations, but is slower in heterogeneous environments.

Urbano et al. [53] use a classical convention emergence game to study the influence of interaction topologies (random, regular, small world and scale-free topologies) on the External Majority strategy update rule (EM) originally proposed by

Shoham and Tennenholtz [47]. In particular, they investigate the effect of different memory sizes on the rate of convergence. Their empirical results show that a small memory that a records a small number of recent interactions is the best option for all types of topologies. More recently, Franks et al. [21] have shown that inserting a small amount of influencer agents — those with specific conventions and strategies — is enough to manipulate the convention adopted by large societies. In a particular effort, Delgado et al. [14,15] study the emergence of coordination in scale-free networks. Their study involves an interaction model of a multi-agent system, by which they analyse how fast coordination can spread among agents. Coordination here is represented through agents being in the same state, which is viewed as being achieved when 90% of the agents are in the same state. The results of the work demonstrate that coordination can indeed be achieved over scale-free networks, but in a rather restricted setting. Similarly, Sen and Sen [45] analyse the effect of increasing the number of actions available to agents, as well as the effect on the speed of norm emergence, of increasing the number of agents. Their results suggest that both increasing the number of actions *and* increasing the number of agents causes a delay to norm emergence in the population.

Villatoro et al. [56] adopt the same concept of two-agent interactions, but introduce the notion of the reward of an action being determined through the use of the memory of agents, thus adding some dynamism to the model. Here, the reward of a certain action is determined by whether the action represents the majority action in both agents' memories, and the reward is proportional to the number of occurrences of this majority action in their memories. However, it is not clear from where these rewards derive nor who applies them, as agents only have access to their memory.

3 Axelrod's Model

In Axelrod's model, a population of agents play a game in which each agent has to decide between cooperation and defection [2]. The agent population evolves through a number of iterations, with a mechanism whereby successful behaviour (as measured by the scoring system) tends to be replicated and unsuccessful behaviour tends to be discarded. In each iteration, each replicated behaviour is subjected to a small chance of mutation, reflecting the feature that an agent may occasionally change its strategy, irrespective of past habits. The strategy of each agent in determining whether to defect and whether to punish others is based on two different attributes, *boldness* (encouraging agents to defect) and *vengefulness* (encouraging them to punish others), which are distinct for each agent. The idea is that a system eventually resulting in all agents having high vengefulness and low boldness corresponds to norm emergence, since they will punish defection but they will not themselves defect. Axelrod's model is introduced in two stages, detailed next.

3.1 The Norms Game

Axelrod's *norms game* adopts an evolutionary approach in which successful strategies are multiplied over generations, potentially leading to convergence of norms.

Term	Description	Value
i, j	Individuals	A numerical index to identify individual agents
S	Probability of a defection being seen by any given individual	Uniform distribution from 0 to 1
B_i	Boldness of i	Uniform distribution from $\frac{0}{7}$ to $\frac{7}{7}$
V_i	Vengefulness of i	Uniform distribution from $\frac{0}{7}$ to $\frac{7}{7}$
T	Player's temptation to defect	$+3$
H	Hurt suffered by others as a result of an agent's defection	-1
P	Cost of being punished	-9
E	Enforcement cost, i.e. cost of applying punishment	-2

Table 1 The Norms Game terms (from [2])

Each individual, or agent, can choose to *defect* by violating a norm, and such behaviour has a particular known chance of being observed or *seen* (S). An agent i has two decisions, or strategy dimensions, as follows. First, it must decide whether to defect, determined by its *boldness* (B_i). Second, if it sees another agent defect (determined by S), it must decide whether to punish this defecting agent, determined by its *vengefulness* (V_i), which is the probability of doing so. (A full list of the terms used in the norms game is provided in Table 1). If $S < B_i$, then agent i defects, receiving a *temptation payoff*, $T = 3$, while *hurting* all the others with payoff $H = -1$. If a defector is *punished* (P), the payoff to the defector is $P = -9$, while the punishing agent pays an *enforcement cost* $E = -2$. The initial values of B_i and V_i are chosen at random from a uniform distribution of a range of eight values between $\frac{0}{7}$ and $\frac{7}{7}$.

Axelrod's simulation has a population of 20 agents, with each agent having four opportunities to defect, and the chance of being seen for each drawn from a uniform distribution between 0 and 1. After playing a full round (all four opportunities), scores for each agent are calculated in order to produce a new generation, as follows. Agents that score better or equal to the average population score plus one standard deviation are reproduced twice in the new generation. Agents that score one standard deviation under the average population score are not reproduced at all, and all others are reproduced once. Although this may produce a new generation with a different number of agents, Axelrod maintains the number of agents at 20 over subsequent generations, but does not specify how. Finally, a mutation operator is used to enable new strategies to arise. Since B_i and V_i (which determine agent behaviour) take eight possible values, they need three bits to be represented. Mutation is thus applied by flipping one of these bits whenever an agent is reproduced, with a 1% chance and a uniform selection of which bit to flip.

In terms of experimental results of the norm game, Axelrod's original experiment [2] has five runs, each with 100 generations. The final results of these runs are as follows: two runs achieved high average boldness and almost zero average vengefulness, indicating no norm emergence at all; two other runs achieved low average boldness and vengefulness; and only the final run achieved a high level of average vengefulness and very low average boldness, indicating the partial establishment of a norm against defection. These results are not satisfactory, since they indicate that norm establishment is not guaranteed. In response, Axelrod intro-

Term	**Description**	**Value**
i, j	Individuals	A numerical index to identify individual agents
S	Probability of a defection being seen by any given individual	Uniform distribution from 0 to 1
B_i	Boldness of i	Uniform distribution from $\frac{0}{7}$ to $\frac{7}{7}$
V_i	Vengefulness of i	Uniform distribution from $\frac{0}{7}$ to $\frac{7}{7}$
T	Player's temptation to defect	$+3$
H	Hurt suffered by others as a result of an agent's defection	-1
P	Cost of being punished	-9
E	Enforcement cost, i.e. cost of applying punishment	-2
$P\,'$	Cost of being punished for not punishing a defection	-9
$E\,'$	Cost of punishing someone for not punishing a defection	-2

Table 2 Metanorms Game terms (from [2])

duces a metanorm model, which incorporates an additional mechanism to better support norm establishment.

Of course, there is no reason why a society in which norms do emerge should be seen as "better" than one in which they do not. Many people argue that deregulation (of, for example, financial markets) is needed if enterprise and innovation is to thrive. It is the existence of metanorms that suggests that norms are indeed desirable. In Section 4 we consider why norms might not emerge without the metanorm.

3.2 The Metanorms Game

The key idea underlying Axelrod's metanorm mechanism is that some further encouragement for enforcing a norm is needed. This is accomplished by introducing a *metanorm* for punishment of those who observe a defection but do not punish the defectors. In this new metanorms game, if an agent sees a defection but does not punish it, this is considered as a different type of defection, and others in turn may observe this defection (with probability S) and apply a punishment to the non-enforcing agent. However, if agents decide not to apply this second level punishment, there is no risk of another level of punishment being applied to them. As before, the decision to punish is based on vengefulness, and brings the defector a punishment cost of $P' = -9$ and the punisher an enforcement cost of $E' = -2$ (see Table 2 for list of terms used in the metanorm game). Applying this new metanorm game to the same simulation as before gives runs with high vengefulness and low boldness, which is exactly the kind of behaviour needed to support establishment of a norm against defection.

Introducing metanorms expresses a preference for norms over no norms, imposes an obligation to punish on agents, rather than relying on individual temperaments, and their acceptance of such a metanorm suggests acceptance of membership of the society, with accompanying social obligations, on the part of the agents.

4 Analysis of Axelrod's Model

It can be concluded from last section that norm establishment is always guaranteed using the metanorms game. However, the experiments undertaken by Axelrod are limited in terms of scale and duration. Binmore [5] has also identified similar concerns in relation to the limits of the simulations undertaken. While he criticises the lack of theory in Axelrod's work as part of a broader analysis, he also recognises that it offers some new conjectures that are worth exploring. Clearly, a more comprehensive analysis is needed, but first we need to replicate Axelrod's original experiments in order to clarify certain assumptions that are not obvious.

4.1 Replicating Axelrod's Results

In seeking to replicate Axelrod's results, some assumptions need to be made, about which Axelrod says nothing. Firstly, the model does not specify how the constant population level is maintained after reproduction, when there are three possible scenarios.

1. The new population is smaller than the original. In our re-implementation, additional agents are randomly selected from the resulting population and replicated. Here, if eliminating poorly performing agents, and replicating agents that perform well, as specified above, results in too few agents in the population, we select further agents at random to replicate, up to the number needed to maintain a constant level.
2. The new population is equal in size to the original, in which case no action is needed.
3. The new population is larger than the original. In this case, our position is to select the required number of agents at random from the relevant set for reproduction. Here, if eliminating poorly performing agents and replicating agents that perform well, as specified above, results in too many agents in the population, we select further agents at random to eliminate, to arrive at the number needed to maintain a constant level.

Secondly, we assume the score of each agent is set to 0 at the beginning of each generation.

Based on the above, and the description in Section 3 of Axelrod's model, we are able to provide an algorithmic representation of the model in Algorithm 1. This algorithm covers both Axelrod's norms game and his metanorms game, where lines 11-16 are relevant only in the context of the addition of the notion of metanorm for the metanorms game. Note that in the algorithm, TS_i refers to the total score of agent i and similarly TS_j refers to that for agent j.

Using the algorithm above, Axelrod's experiments were repeated, running the norms game 10 times. The results are shown in Figure 1, where the diamonds represent the value of the mean average boldness and vengefulness of the final generation's population. As can be concluded from the figure, the results obtained are similar to those of Axelrod, with one run having high vengefulness and low boldness, two runs with exactly the opposite (high boldness and low vengefulness), and all other runs with low values for both boldness and vengefulness.

Algorithm 1: The Simulation Control Loop: $simulation(T, H, P, E)$

```
for each agent $i$ do
  for each opportunity to defect $o$ do
    if $B_i > S_o$ then
      $TS_i = TS_i + T$
      for each agent $j : j \neq i$ do
        $TS_j = TS_j + H$
        if see($j$,$i$,$S_o$) then
          if punish $(j, i, V_j)$ then
            $TS_i = TS_i + P$
            $TS_j = TS_j + E$
          else
            for each agent $k : k \neq i \wedge k \neq j$ do
              if see($k$,$j$,$S_o$) then
                if punish $(k, j, V_k)$ then
                  $TS_k = TS_k + E$
                  $TS_j = TS_j + P$
$AvgS$ = calcAverageScore()
$StdD$ = calcStandardDeviation()
for each agent $i$ do
  if $TS_i > AvgS + StdD$ then
    replicate()
  else
    if $TS_i < AvgS - StdD$ then
      eliminate()
maintainPopSize()
```

In order to establish how these results arise, changes to boldness and vengefulness for each individual were monitored. Figure 2 shows the results for three sample individuals, illustrating how the average boldness and vengefulness vary over generations (and hence time). In particular, Figure 2(a) shows one run ending in the most common result, low boldness and low vengefulness. The run starts with average boldness and vengefulness of about 0.5 (as initial values for B_i and V_i are taken from a uniform distribution over $\{\frac{0}{7}, \ldots, \frac{7}{7}\}$). In the early stages,

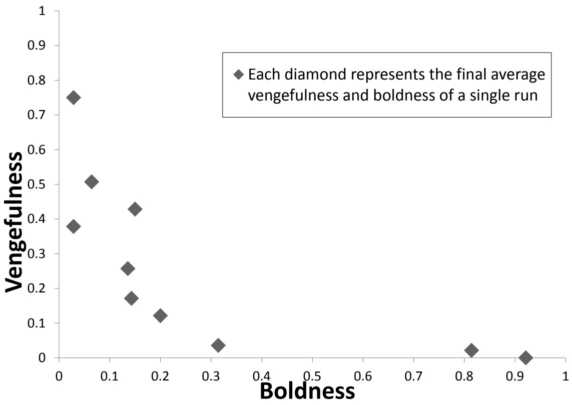
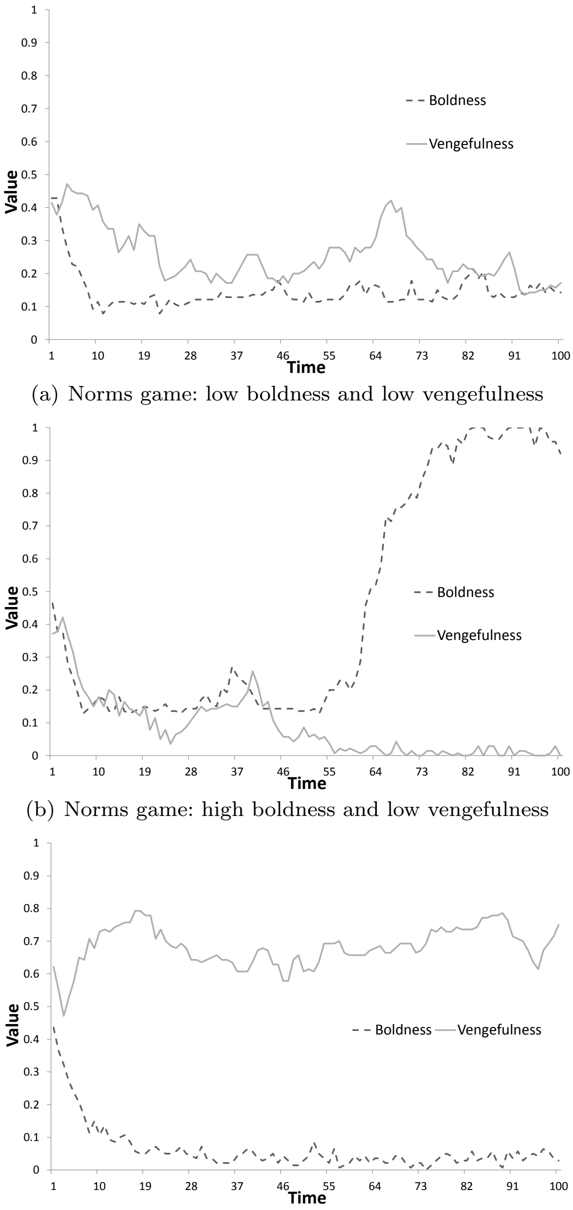


Fig. 1 Norms game overall results



(a) Norms game: low boldness and low vengefulness

(b) Norms game: high boldness and low vengefulness

(c) Norms game: low boldness and high vengefulness

Fig. 2 Norms game: analysis of sample runs

boldness decreases slightly, indicating that individuals with higher boldness are eliminated. This is because high boldness causes an agent to defect, yet defecting in the presence of others with average vengefulness can be costly, as the agent is likely to be punished, leading to a low score. Subsequently, boldness stabilises at a low level. With a low average boldness in the population, being vengeful becomes costly, causing agents with low vengefulness to be favoured over those with high vengefulness, with the latter getting eliminated when forming the new generation. Finally, the values stabilise at particular low values for both boldness and vengefulness until the end of the run.

In the cases that result in high boldness and low vengefulness (an example run is shown in Figure 2(b)), the run starts as before, with both values reducing. However, around the 60th generation, the value of boldness increases sharply until it reaches 1, where it remains. This can be explained by a dramatic change to one individual's boldness, due to mutation, in an agent population with particularly low values of vengefulness. In turn, this facilitates the individual's survival, dominating the others and allowing it to propagate its high boldness across the population. Here, a high score is attained by defecting without punishment (due to low vengefulness), which also hurts others and lowers their scores. In the final case, as shown in Figure 2(c), the run ends with high vengefulness and very low boldness: while eliminating all those with high boldness, only individuals with high vengefulness remain, so there are no individuals with low boldness and low vengefulness, and those with high vengefulness and low boldness survive and dominate.

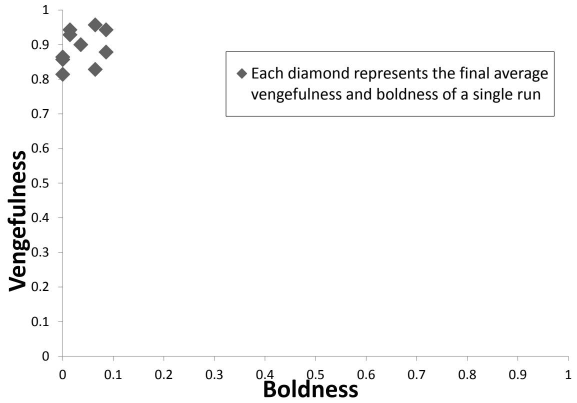


Fig. 3 Metanorms game overall results

By introducing metanorms, Axelrod aimed to address the problems identified above. Our replication of the metanorms game also provides similar results to Axelrod (see Figure 3). Again, a deeper analysis of the results is helpful. As shown in Figure 4, in the metanorms game, the population starts eliminating high boldness individuals as before, but now also eliminates low vengefulness individuals. The latter trend is due to the introduction of the metanorm according to which failure to penalise a defector may also be penalised. This results in a population with high vengefulness and low boldness, which survives until the end of the run.

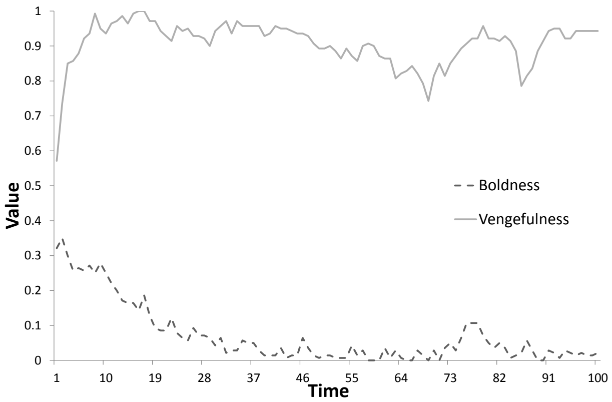


Fig. 4 Metanorms game analysis

In essence, these simulations characterise a situation reminiscent of the Wild West in which individual citizens are relied upon to enforce the law, explaining the importance of vengefulness in the emergence of norms. A"live and let live" attitude is not conducive to maintaining a law, but the burden of punishment here is unevenly spread. However, as these results show, emergence of a norm in such situations is by no means certain. It is this that leads to a requirement for metanorms if norms are to emerge with any certainty.

Such metanorms reflect a shift in attitude on the part of society towards the norms: whereas without metanorms, punishment is seen as a (visceral) response to defection, depending on the vengeful emotions of the victims, the metanorms make it the duty of members of society to punish, so that failure to perform this duty is itself liable to punishment. This move from enforcement by vigilantes (those taking the law into their own hands) to seeing law enforcement as the social duty of responsible citizens is an important milestone in the development of a society that respects its laws. Without this shift, respect for law is hardly even a social virtue, let alone a (social) duty. It is little more than a matter a taste: some agents have a taste for vengeance and some do not. If it is no more than a preference, a matter of taste, then that laws do not always emerge, or if they do, collapse over time, becomes unsurprising. Obedience to laws may or not be to the taste of the society.

4.2 Game Duration

The settings under which Axelrod's experiments are performed are limited, particularly with relation to the duration of these experiments. To provide a stronger analysis of Axelrod's model, the experiments were repeated over a *long duration*, 1,000,000 generations (and 10 runs), as opposed to 100 generations (in Axelrod's experiments and our experiments of Section 4). In our norms game simulation, the game starts with boldness decreasing, and then vengefulness decreasing until they both settle at a low level, which is consistent with Axelrod's results. However, an agent with high boldness can be introduced to such a population through mutation, and would dominate others since it is not punished due to low levels of vengefulness. Clearly, running the experiment for a longer period increases the possibilities for this to occur and, as shown in Figure 5, this always leads to norm collapse.

In undertaking their own analysis of Axelrod's metanorms game, Galan and Izquierdo [22] increased the number of generations in a run and found different results. By including 1,000,000 generations, in 1,000 runs, nearly 70% ended in *norm collapse*, as opposed to Axelrod's finding of *norm establishment*. According to Galan and Izquierdo, this is because vengefulness is costly in a population in which violation is rare. Thus, agents with low vengefulness are favoured over agents with high vengefulness, leading to a significant decrease in vengefulness, encouraging defection, and in turn causing boldness to increase. The results of Galan and Izquierdo suggest that metanorms are not as useful as it might initially seem from Axelrod's results.

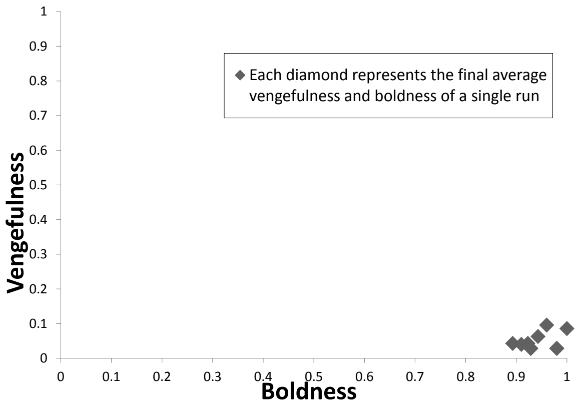


Fig. 5 Norms game for 1,000,000 generations

By analysing these cases to determine the reasons for these results, it is clear that the runs begin in the same way as previously observed, by eliminating individuals with high boldness and low vengefulness, stabilising on those with high vengefulness and low boldness. Then, however, mutation causes vengefulness to reduce. If an agent x with high vengefulness and low boldness changes through mutation to give lower vengefulness, while boldness for all remains low, there is no defection and the mutated agent survives. In addition, if boldness then mutates to become just a little higher for a different agent y, with average vengefulness remaining high, x will still rarely defect because of relatively low boldness. If it *does* defect, and *is* seen by others, it receives a low score, unless it is not punished, in which case the non-punishing agents may themselves be punished because of the high vengefulness in the general population. Here, agent x may not punish y either because of the low probability of being seen (which must be below the low boldness level to have caused a defection) or because it has mutated to have lower vengefulness. In the former case, x will not be punished for non-punishment since it has not observed y's defection, but in the latter case, x might be punished if it is seen by others. However, we know that the probability of being seen is low because agent y has defected (and $S < B$ for defection to take place). In this case, y is eliminated, while x remains, because the likelihood of y's defection being seen by just one agent is relatively high, but the likelihood of agent x's non-punishment being seen requires first y's defection being seen by x, and then x's non-punishment being seen by others, the combination of these being extremely unlikely.

If the values of vengefulness continue to decrease in this way, the population can arrive at a situation with very low average boldness and vengefulness. At this point, a single mutation to boldness could then cause the mutant to dominate the others due to the general lack of vengefulness in the population. The key question here is why, in cases of high boldness and low vengefulness, mutation of vengefulness from a very low value to a significantly higher value does not cause boldness to decrease. Here, such a mutant should punish all others for defecting *and* for not punishing defectors. However, these punishments also incur significant enforcement costs, all of which are borne by the punishing agent, potentially exceeding the penalty meted out to the defectors and those agents who fail to punish others. For such a mutant to survive, it needs to co-exist with other vengeful agents, with which it shares the substantial enforcement load. In other words, there need to be sufficient vengeful agents in the population to overcome this.

This analysis suggests that both mutation and sanctioning costs play a major role in collapsing norms. However, there is an additional factor that gives rise to these results, a particular characteristic of the underlying reproduction policy in the model which, in certain circumstances, and with a very subtle change, can give a very different outcome. We consider this next, in Section 4.3.

We note that there have been similar criticisms of Axelrod's work. In particular, Binmore et al. [6] considered a simplified version of the *ultimatum game* that is also a simplified version of the norms game, concluding that the convergence process is not robust and can easily be disrupted by quite small perturbations, giving little confidence in conclusions obtained from just a few runs of a computer simulation. Indeed, as we have done, Binmore [5] suggests that one needs to conduct very large numbers of tests before beginning to take any results seriously.

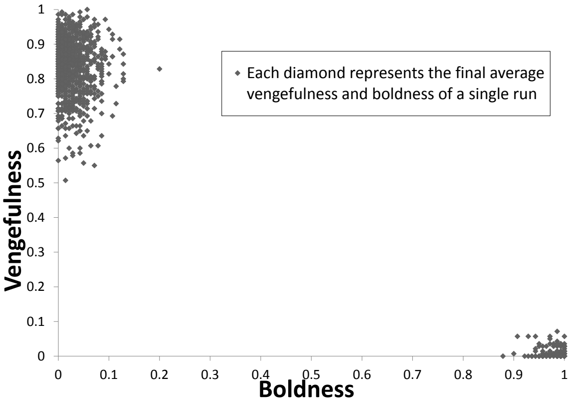


Fig. 6 Metanorms game for 1,000,000 generations and 1000 runs

4.3 Reproduction and Norm Collapse

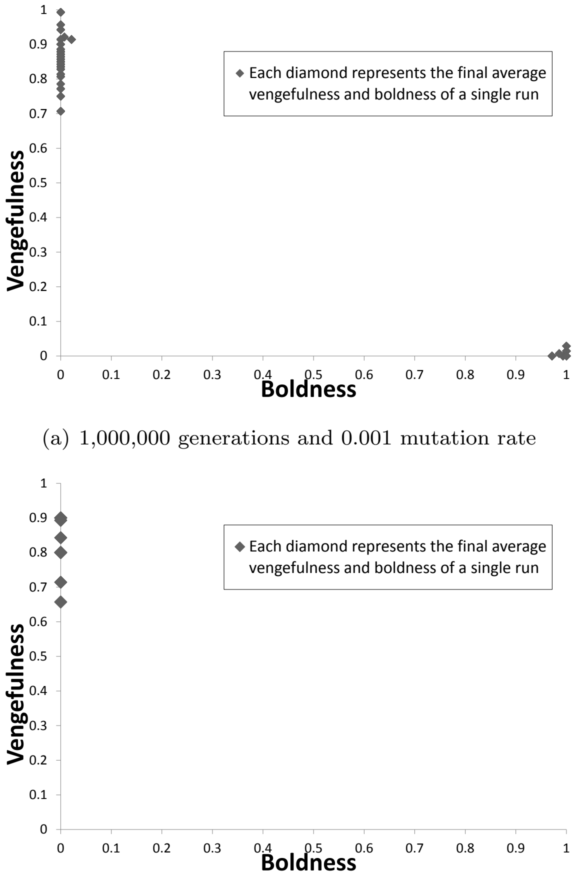
As specified earlier, a run of the metanorm game settles at very low boldness and very high vengefulness at a certain point. For this to change to the opposite situation of low vengefulness and high boldness, a sequence of modifications that lower vengefulness must occur, and another sequence of modifications that increase boldness must also occur.

Another factor that could contribute to norm collapse, which is not considered above, is the reproduction policy over generations, especially where the boldness of the entire agent population is very low. Specifically, when all individuals have boldness around 0, defection rarely happens, and their scores (which change only when agents enforce, are hurt, or are punished) are 0. As a result, the average score and standard deviation are also 0, so that all agents have a score equal to the average score plus one standard deviation. According to Axelrod's rules, agents in this situation should be replicated *twice* when forming the new generation, giving them a better chance of surviving the next round. However, duplicating the individuals in this case does not seem sensible since all of the agents in the population have performed equally, so there is no need to replicate every individual twice. Hence, to study the effect of the reproduction policy, we undertook simulations in which agents are only replicated once in the above special case (as opposed to Galan and Izquierdo where all agents in such a case are duplicated). The results obtained by running the metanorm game with many more generations are similar to those of Galan and Izquierdo, but with a different proportion giving rise to defection. As shown in Figure 6, 128 out of 1,000 runs (or 13%) of 1,000,000 generations ended in norm collapse, as opposed to 70% reported by Galan and Izquierdo.

The reason behind the different results is that replicating an entire population of non-defecting agents increases the likelihood of significant fluctuations in vengefulness over subsequent generations. For example, in one phase of a run using the approach of Galan and Izquierdo in which all agents have 0.0 boldness, 5 agents have vengefulness of 0.0, 11 with 1.0 and 4 with 0.8, the next generation includes 8 agents with vengefulness of 0.0, 7 agents with 1.0, and 5 agents with 0.8, simply due to the replication policy. This means that average vengefulness drops from 0.71 to 0.55 and, as boldness continues at 0.0, replication again makes this worse. However, note that replication could cause the opposite, increasing the number of agents with high vengefulness over those with low vengefulness.

As new generations of agents with low boldness are evolved for more iterations, it becomes more likely to observe the following combination of events. First, the levels of vengefulness decrease repeatedly through a sequence of downward fluctuations until they reach very low levels. Then, until vengefulness levels fluctuate back upwards, this creates a temporarily fertile environment for defectors. Next, the boldness of one or a few agents increases to a high level due to mutation. This causes the bold agents to defect and be replicated in subsequent iterations of the game. The end result of the phenomenon is an agent population where high boldness and defection are so prevalent that being vengeful leads to extinction. Thus, the game reaches a stable situation where norm collapse is ingrained in the population (i.e. low vengefulness and high boldness).

This end state is reached in a proportion of experimental runs of both Galan and Izquierdo's experiments and our own experiments because it is more likely to reach the required preconditions when repeating the experiment for 1,000,000



(a) 1,000,000 generations and 0.001 mutation rate

(b) 1,000,000 generations and 0.0001 mutation rate

Fig. 7 Metanorms game with different mutation rates

generations (i.e. over a long duration). However, a much larger proportion of Galan and Izquierdo's runs end in stable norm collapse because their replication policy allows for much more significant fluctuations of vengefulness.

4.4 Mutation

As discussed above, mutation is significant in determining when norm collapse occurs. Galan and Izquierdo [22] argue that decreasing the mutation rate from 0.01 to 0.001 allows norm collapse to arise much earlier. They present an example in which a mutation rate of 0.001 allows the population to converge towards norm collapse in about 25,000 generations as opposed to 300,000 with a rate of 0.01. However, they do not explain the reasons, and do not explore the more general effect of mutation rate on norm collapse or establishment.

In seeking to consider this further, we undertook an experiment consisting of 1000 runs of 1,000,000 generations, using the mutation rate of 0.001 suggested by Galan and Izquierdo, with results shown in Figure 7(a). Clearly, these results support Galan and Izquierdo's argument: 40% of the runs ended in norm collapse. However, decreasing the mutation rate further does not have the same effect. In particular, a mutation rate of 0.0001 resulted in all 1000 runs ending in norm establishment, shown in Figure 7(b).

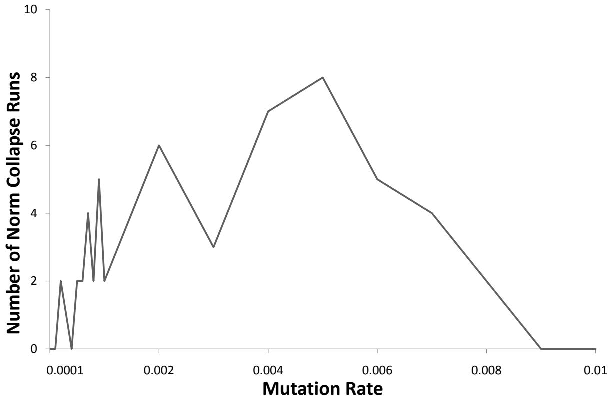


Fig. 8 Metanorms game: 1,000,000 generations; 0.0001–0.01 mutation rate

The relation between mutation rate and norm collapse is thus unclear. To understand this better, we performed a further series of experiments. Figure 8 illustrates the result of different experiments that consists of 10 runs each, with a range of mutation rates between 0.0001 and 0.01. As can be observed from Figure 8, the mutation rate seems to play an important role in causing norms to collapse. Decreasing the mutation rate below 0.01 has a major effect on the proportion of runs ending in norm collapse, with a peak around mutation rate values of 0.005 giving norm collapse in 80% of runs. However, decreasing the mutation rate further causes the proportion of runs ending in norm collapse to drop back (with fluctuations) until it reaches 0 with a mutation rate of 0.0001. While these results suggest a potentially interesting relationship, further work is needed to establish the exact correlation. Since the focus of this paper is to examine the possibility of applying metanorms to computational systems, and it is difficult to justify the use of mutation in these systems, this is outside scope but remains an important area for future work.

Nevertheless, we can say that given these results, removing mutation should avoid norm collapse. In the norms game, after the population stabilises at a low level of both vengefulness and boldness, mutation of an agent's boldness from low to high allows it to dominate, and as a result eliminate others, which leads to norm collapse (as previously shown in Figure 5). Figure 9 illustrates the result of a no-mutation norms game that consists of 1000 runs, with 1,000,000 generations each. As expected, removing mutation avoids norm collapse and leaves the population with the other two situations.

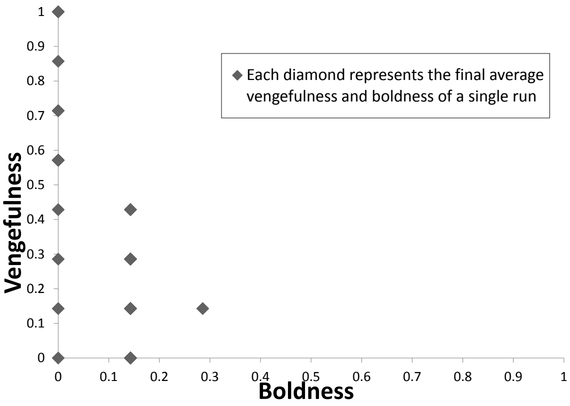


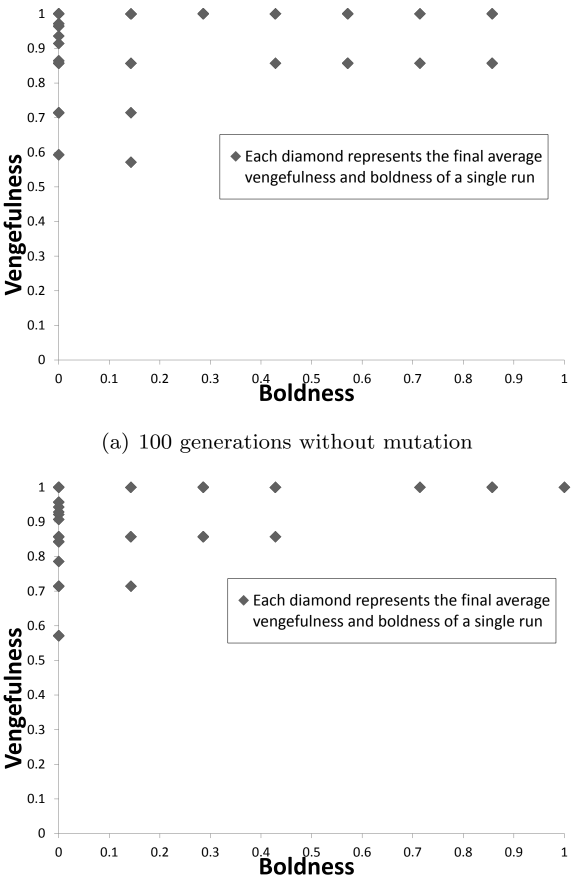
Fig. 9 No-mutation norms game: 1,000,000 generations and 1000 runs

In the metanorms game, as we have seen, mutation seems to have a great effect on moving away from norm establishment. By removing mutation, we might expect to guarantee norm establishment. To corroborate this, we performed two experiments for two different durations, with results shown in Figure 10(a) for 100 generations and 1000 runs and Figure 10(b) for 1,000,000 generations and 1000 runs. Surprisingly, the results were not as expected. A high level of vengefulness is maintained in almost all the runs, but a high level of boldness is also observed in some, and hence a high level of defection in the population, despite the associated punishment. This is because the final result of each run primarily depends on the initial distribution of vengefulness and boldness: if all individuals with high vengefulness happen to have high boldness, and if those with mid or low vengefulness have low boldness, then individuals with high vengefulness and high boldness at the start are favoured over those with average vengefulness and low boldness. As a result, they survive and dominate the others. More importantly, the final result is determined within the very first generation so that running the experiment for a longer period has no impact and there is no change to the population once the levels of vengefulness and boldness stabilise.

4.5 Characterising the vengefulness-boldness space

It is clear that Axelrod's model exhibits many interesting aspects, and relies on characteristics that provide different results with different assumptions or instantiations. We have explored several of these in relation to our experiments, found some distinct features and, as a result, we can also provide a characterisation of the nature of norm establishment or collapse more generally.

Given the analysis so far, it should be clear that norm establishment lies in the region where vengefulness is high and boldness is low; similarly, norm collapse lies in the region where vengefulness is low and boldness is high. This is as used by Axelrod in his model, and underlies the aim of the initial experiments. It is illustrated graphically in Figure 11. However, we can also characterise the region where vengefulness is low and boldness is low as tending to norm collapse: this is



(a) 100 generations without mutation

(b) 1,000,000 generations without mutation

Fig. 10 No-mutation metanorms game

a region of benign behaviour since boldness is low and defection is unlikely, but it is unstable since a mutation to boldness may take it higher, leaving vengefulness low, and causing norm collapse. Conversely, the region where both vengefulness and boldness are high is tending to norm establishment: it is undesirable since boldness is high and there are many defections, but these defections are likely to be punished. If vengefulness does cause punishment, then it is likely that boldness will drop, leading to norm establishment. Given this view, we can reinterpret the previous experiments. For example, it is clear that the run shown in Figure 2(a) ends in a state tending to norm collapse, the run in Figure 2(b) ends in norm collapse, and the run in Figure 2(c) ends in norm establishment, as does the run in Figure 3.

From the results we have seen, it is clear that a lack of vengefulness can lead to norm collapse when a bold mutant arrives. In the real world, this is often the point at which we sometimes see the emergence of a central authority: in recognising the threat to order presented by the arrival of bold deviant individuals,

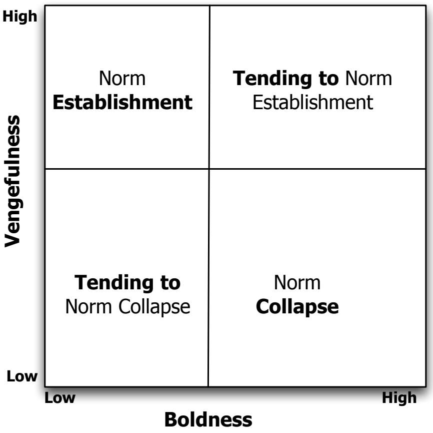


Fig. 11 Characterising the vengefulness-boldness space

and being themselves reluctant to punish because of the cost involved, the citizens tend to find a different mechanism for punishing, and hence suppressing, the bold individuals. Here, what they do is pay a levy to a central authority, which punishes the defectors. If there are relatively few law breakers (and where there is in general low boldness), a relatively small levy will cover the costs of punishment (and possibly allow the lawman a profit). Returning to the Wild West analogy considered earlier, as communities developed, the citizens would club together to hire a gunfighter to keep the peace. Alternatives include seeing this as insurance, or possibly establishing a central fund out of which agents who punish are rewarded (which we might see as analogous to bounty hunters). Of course, if there are too many defectors the levy will become too high, but given a low boldness and low vengeance society, the levy should be advantageous over enforcement costs. (This explains why the central authority will not emerge until the level of boldness has fallen to a relatively low level.) The essential point is that norm establishment requires the punishment of bold individuals who defect, and if the citizens are not themselves vengeful enough, they need to pay someone to do it for them. Note also that it is expected and desired that defectors be punished, since the agents have already incurred the cost of punishment. While the process can be seen as the formation of a Hobbesian social contract, there remain problems with this centralised perspective, not least among them how to detect violations and violators. Indeed in this paper, our focus is on the emergence of norms rather than a centralised mechanism for enforcement, so while we note the alternatives, we omit further discussion of such possibilities and leave them for further consideration outside the context of this work.

5 Omniscience in Axelrod's Metanorm Model

From the above analysis, it is clear that various improvements are needed in order to generalise the metanorms approach and make it applicable over computational systems. First, the model assumes full control over the entire agent population, which permits access to all agent utility scores and allows central manipulation of the entire population by adding or eliminating agents. However, this is rarely possible in computational systems in which populations can be vast, and no single party can have control over every agent. Furthermore, agent scores are typically understood as private information that agents might choose not to share with others, posing a further problem for the application of Axelrod's approach.

Instead, we need a mechanism through which individuals can learn to improve their strategies over time. If we enable individuals to compare themselves to others, and adopt more successful strategies, then we can take a *learning interpretation* of the evolutionary mechanism [39], without needing to remove and replicate individuals. However, this learning interpretation requires that the private strategies of individuals are available for observation by other agents, which is again an unreasonable assumption. As we have shown in the previous sections, Axelrod's model is not capable of sustaining cooperation over a large number of generations. Furthermore, Axelrod's approach relies on agents being able to punish both those that defect and those that fail to punish defection, yet this is unrealistic since it assumes *omniscience* through agents being aware of all norm violations and punishments.

In response, the remainder of this paper investigates alternatives that allow us to make use of the mechanisms resulting from Axelrod's investigations, in more realistic settings. Specifically, we first take a learning interpretation of evolution and describe an alternative technique, strategy copying, which prevents norm collapse in the long term. Second, we remove the assumption of omniscience and constrain the ability of agents to punish according to the defections they have observed. Finally, to obviate the need for knowledge of the private strategies of others, we propose a learning algorithm through which individuals improve their strategies based on their experience.

5.1 Strategy Copying

In Axelrod's model, the evolutionary approach causes some problems in extended runs, leading to norm collapse. In addition, for use in domains such as peer-to-peer

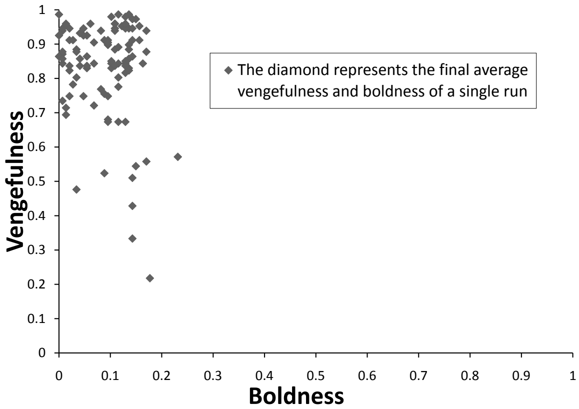


Fig. 12 Strategy copying from the best agent, 100 timesteps

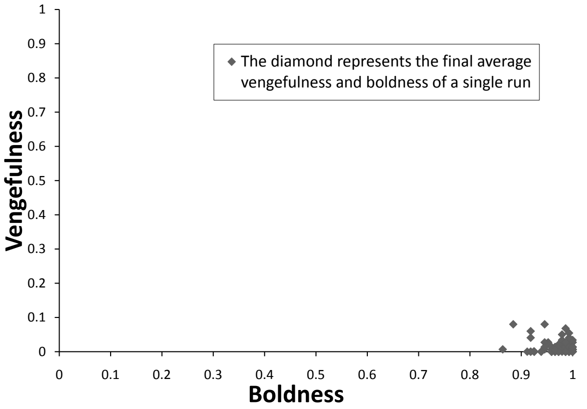


Fig. 13 Strategy copying from the best agent, 1,000,000 timesteps

or wireless sensor networks, the agents themselves cannot be deleted or replicated, but instead must modify their own behaviour. In this section, therefore, we examine a simple alternative to Axelrod's model in which an agent that performs poorly in comparison to others in the population can *learn* new strategies (in terms of vengefulness and boldness attributes) by adopting the strategy of other, better performing agents, replacing the existing strategy with a new one. Agents can achieve this in different ways: they can copy the strategy of the agent with the highest score or they can copy the strategy of one in a group of agents performing best in the population. It is important to note that the parameter set-up used in all experiments introduced in this section is the same as that specified in Table 2.

5.1.1 Strategy Copying from a Single Agent

Intuitively, copying the strategy of the agent with the highest score appears to be a promising approach. However, it leads to poor results in the long term because it draws on the strategy of a single agent rather than a population of agents. This makes the approach vulnerable to strategies that are only successful in a small number of possible settings. Moreover, by failing to draw on strategies from a variety of agents, the strategies tend to converge prematurely.

To illustrate this, consider a group of students taking an examination, with one of the students having cheated. If the cheating student has not been seen, they may achieve the best exam performance. However, if all other students copy this behaviour and cheat in the next exam, there is a high possibility that they will be caught, and will thus suffer from much worse results than if they had not cheated. This is supported by the results shown in Figures 12 and 13, illustrating experiments with runs of 100 and 1,000,000 timesteps (where a timestep represents one *round* of agents having opportunities to defect and learning from the results, and is equivalent to a *generation* in the evolutionary approach). Each point on the graph represents the average boldness and vengefulness of the population at the end of a single simulation run.

In the short term, as can be seen from Figure 12, copying from the best agent leads to norm establishment. However, in the long term the norm collapses, as

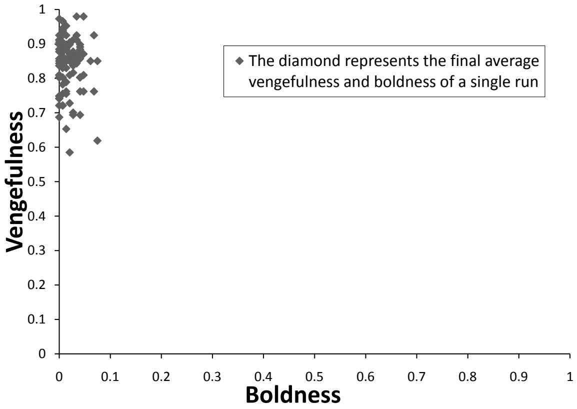


Fig. 14 Strategy copying from a group of agents, 1,000,000 timesteps

shown in Figure 13. This can be explained by the fact that an agent with low vengefulness that does not punish a defector (and thus does not pay an enforcement cost) but is also not metapunished, scores better than any other agent with high vengefulness that *does* punish (and thus pays the enforcement cost). As a result, other agents copy the low vengefulness of this agent so that low vengefulness becomes prevalent in the population. In the same way, when low vengefulness prevails in the population, an agent with high boldness defects, gaining a *temptation payoff*, and hurting others without receiving punishment. As a result, other agents copy the high boldness of this agent so that low vengefulness and high boldness is propagated through the population, leading to norm collapse. The essence of the free-rider problem is that deviation is highly profitable to the free-riders as long as sufficient numbers continue to conform, but that if everyone tries to free-ride, all benefits are lost (c.f. the Nash Equilibrium strategy of mutual defection in the Prisoners? Dilemma).

5.1.2 Strategy Copying from a Group of Agents

Alternatively, as we have suggested, we might seek to copy the strategy of one in a group of high-performing agents. In this view, agents choose one agent, at random, from the group of agents with scores above the average, and copy its strategy. As previously, experiments of different durations (between 100 and 1,000,000 timesteps) were carried out. The results in Figure 14, for 1,000,000 timesteps, show that all runs ended with norm establishment in the long term, indicating that this approach is effective in eliminating the problematic effect of the replication method. This approach avoids norm collapse since it does not limit itself to the best performing agent, and thus does not run the risk of only adopting a strategy that performs well in a small number of settings.

5.1.3 Observation of Defection

In Axelrod's model (see Section 3), an agent z is able to punish another agent y that does not punish a defector x, even though agent z does not see the defection of

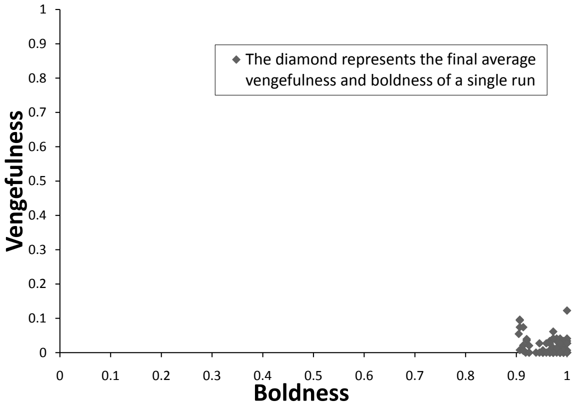


Fig. 15 Strategy copying with defection observation constraint, 1,000,000 timesteps

agent x. However, such metapunishment is not possible if the original defection is not observed: guaranteed observation of the original defection is an unreasonable expectation in real-world settings. In consequence, our model needs adjustment so that metapunishment is only permitted if an agent observes the original defection. However, because this observation constraint limits the circumstances in which metapunishment is possible, its introduction corresponds to removing the metapunishment component from part of the game. In Axelrod's original experiments, metapunishment was introduced as a means to stabilise an established norm, since norms tend to collapse shortly after they are established without metapunishment. In fact, this remains the case in our model and our results confirm this.

More precisely, the observation constraint causes all runs to end in norm collapse when simulations are run for 1,000,000 timesteps, as shown in Figure 15. This is due to the fact that, as in the original model, runs initially stabilise on high vengefulness and low boldness, and then mutation causes vengefulness to reduce. If an agent y with high vengefulness and low boldness changes through mutation to give lower vengefulness, while boldness for all remains low, there is no defection and the mutated agent survives. If boldness then mutates to become just a little higher for a different agent x, with average vengefulness remaining high, x will still rarely defect because of relatively low boldness.

If it *does* defect, however, and *is* seen by others, it receives a low score, unless it is *not* punished, in which case the non-punishing agents may themselves be punished because of the high vengefulness in the general population. Here, agent y may not punish x because of the low probability of being seen (which must be below the low boldness level to have caused a defection) or because it has mutated to have lower vengefulness. In the former case, y will not be metapunished for non-punishment, but in the latter case, y might be metapunished if it is seen by others. The likelihood of agent y's non-punishment being seen requires first x's defection being seen by y, and then y's non-punishment being seen by others. Importantly, in this new model, agents that metapunish y must themselves see x's defection. Since this combination of requirements is rare, such mutants like y survive for a longer duration, enabling their strategy to propagate through the population, and causing vengefulness to decrease. In addition, if another such event occurs, it will cause the average vengefulness of the population to drop further until it reaches a very low level. When the model runs over an extended period, such a sequence of events is much more likely, and low vengefulness allows a mutant of higher boldness to survive and spread among the whole population, which is the cause of the norm collapse. Indeed, where agents have subscribed to a central enforcement agency, it will be in their interests to report any defections they observe, since they have already expressed support for the norms, and have already incurred the cost of punishment.

While our focus is on these aspects of Axelrod's original model, we note that an alternative approach could be taken, in which the metapunisher's knowledge of agents' capability of observing is the key driver. This would allow metapunishers to punish independently of whether other agents have in fact observed violations (and this does sometimes happen in the law, where a metapunisher considers if it is reasonable to assume that another agent should presume that some wrongdoing was committed). Indeed, this would be a valuable aspect to explore, but would take the paper in a different direction, away from our desired relative fidelity to Axelrod, and is beyond the scope of our work.

5.2 Strategy Improvement

Once the observation constraint is introduced, strategy copying becomes inadequate and leads to norm collapse. Furthermore, it requires that agents have access to the strategies and decision outcomes of others in order to enable the copying mechanism. As we have already argued, in real-world settings such observations tend to be unrealistic. *Reinforcement learning* offers an alternative to Axelrod's evolutionary approach to improving performance of the society while keeping agent strategies and decision outcomes private. There are many reinforcement techniques in the literature, such as Q-learning [58], Policy Hill Climbing (PHC) [10] and WOLF-PHC [10], which we use as inspiration in developing a learning algorithm for strategy improvement in the metanorms game.

5.2.1 Q-learning

Q-learning is a reinforcement learning technique that allows the learner to use the (positive or negative) reward, gained from taking a certain action in a certain state, in deciding which action to take in the future in the same state. Here, the learner keeps track of a table of Q-values that records an action's quality in a particular state, and updates the corresponding Q-value for that state after each action. The new value is a function of the old Q-value, the reward received, and a learning rate, δ, and the action with the highest updated Q-value for the current state is chosen. However, for us, Q-learning suffers from two drawbacks. First, it considers an agent's past decisions and corresponding rewards, which are not relevant here; doing so would inhibit an agent's ability to adapt to new circumstances. Second, actions are precisely determined by the Q-value; there is no probability of action, unlike Axelrod's model.

Bowling and Veloso [10] proposed policy hill climbing (PHC), an extension of Q-learning that addresses this latter limitation. In PHC, each action has a probability of execution in a certain state, determining whether to take the action. Here,

Table 3 Effects of actions on agent score

Decision	Effects
Defect	Gain temptation payoff Hurt all other agents Potentially suffer punishment cost
Cooperate	—
Punish	Pay enforcement cost
Not punish	Potentially suffer metapunishment cost
Metapunish	Pay enforcement cost
Not metapunish	—

the probability of the action with the highest Q-value is increased according to a learning rate δ, while the probabilities of all other actions are decreased in a way that maintains the probability distribution, with each probability update occurring immediately after the action. In enhancing the algorithm, a *variable* learning rate is introduced, which changes according to whether the learner is winning or losing, inspired by the WOLF technique (win or learn fast). This suggests two possible values for δ: a low one to be used while an agent is performing well and a high one to be used while the agent is performing poorly.

However, in one round of Axelrod's game, an agent can perform multiple punishments (potentially one per defection and non-punishment observed), while only having a small number of opportunities to defect (four in Axelrod's configuration). Therefore, punishment and metapunishment actions would be considered much more frequently than defection, leading to disproportionate update of probabilities of actions, with some converging more quickly than others. To address this imbalance, we can restrict learning updates to occur only at the end of each round, rather than after each individual action, so that boldness and vengefulness are reconsidered once in each round and evolve at the same speed. The aim here is to change the probability of action significantly when losing, while changing it much less when winning, providing more opportunities to adapt to good performance.

In summary, while basic Q-learning is not appropriate because of the lack of action probabilities, WOLF-PHC suffers from a disproportionate update of such probabilities in this setting. Nevertheless, the use of the variable learning rate approach in WOLF-PHC is valuable in providing a means of updating the boldness and vengefulness values in determining which action to take. However, since agents that perform well need not change strategy, we can consider only one learning rate. The next section details our algorithm, inspired by this approach.

5.2.2 BV Learning

To address the concerns raised above, in this section, we introduce our BV (Boldness-Vengefulness) learning algorithm. This requires an understanding of the relevant agent actions and their effect on the agent's score, as summarised in Table 3, which outlines the different actions available to an agent and the consequence of each on the agent's score.

Now, since boldness is responsible for defection, an agent that obtains a good score as a result of defecting should increase its boldness, and an agent that finds defection detrimental to its performance should decrease its boldness. Learning

Algorithm 2: The Simulation Control Loop: $simulation(T, H, P, E, \gamma, \delta)$

```
for each round do
    interact($T$, $H$, $P$, $E$)
    learn($\gamma$, $\delta$)
```

suitable values for vengefulness is more complicated, since while it is responsible for both punishment and metapunishment, these also cause enforcement costs that decrease an agent's score. Low vengefulness allows an agent to avoid paying an enforcement cost, but can result in receiving metapunishment. Vengefulness thus requires a consideration of all these aspects.

First, in order to determine the unique effect of each individual action on agent performance, we decompose the single combined total score (TS) of the original model into distinct components, each reflecting the effect of different classes of actions. Therefore, each agent keeps track of four different utility values: the *defection score* (DS) of an agent who defects, which is composed of the positive temptation value and the negative punishment incurred, the *punishment score* (PS) incurred by an agent who punishes or metapunishes another (as a result of an enforcement cost), and the *punishment omission score* (POS) incurred by an agent who does not punish another when it should, and is consequently metapunished. These are combined into a total score (TS).

In this context, we can consider the algorithms used in our simulation, in two phases, as represented in Algorithms 3 and 4, which are called by Algorithm 2. (Note that we use subscripts to indicate the relevant agent only when needed). More precisely, in Algorithm 3, each agent has various defection opportunities (o), and defects if its boldness is greater than the probability of its defection being seen (S_o). If an agent defects (Line 3), its DS increases by a *temptation payoff*, T (Line 4), but it *hurts* all others in the population, whose scores decrease by H (line 6), where H is a negative number that represents the hurt suffered. If an agent cooperates, no scores change. DS thus determines whether an agent should increase or decrease boldness in relation to its utility.

Each hurt agent can in turn observe the defection and react to it with punishment probabilistically according to its vengefulness. Punishment and metapunishment both have two-sided consequences: if an agent j sees agent i defect in one of its opportunities (o) to do so, with probability S_o (Line 7), and decides to punish it, which it does with probability V_j (Line 8), i incurs a punishment cost, P, to its DS (Line 9), while the punishing agent incurs an enforcement cost, E, to its PS (Line 10). If j does not punish i, and another agent k sees this in the same way as previously (Line 13), and decides to metapunish (Line 14), then k incurs an enforcement cost, E, to its PS, and j incurs a punishment cost P to its POS. Note that both P and E are negative values, so they are added to the total to reflect their effect.

In Axelrod's original model, those agents that are one standard deviation or more below the mean are eliminated and replaced in the subsequent population generation with new agents following the strategy captured by the boldness and vengefulness values of those agents that are one standard deviation or more above the mean. Thus, poorly performing agents are replaced by those that perform much better. In contrast, in our model, in Algorithm 4, we simply distinguish be-

Algorithm 3: interact(T, H, P, E)

```
for each agent $i$ do
  for each opportunity to defect $o$ do
    if $B_i > S_o$ then
      $DS_i = DS_i + T$
      for each agent $j : j \neq i$ do
        $TS_j = TS_j + H$
        if see($j$,$i$,$S_o$) then
          if punish ($j, i, V_j$) then
            $DS_i = DS_i + P$
            $PS_j = PS_j + E$
          else
            for each agent $k : k \neq i \wedge k \neq j$ do
              if see($k$,$j$,$S_o$) then
                if punish ($k, j, V_k$) then
                  $PS_k = PS_k + E$
                  $POS_j = POS_j + P$
```

tween good and poor performance, with only agents that score below the mean reconsidering their strategy. Thus, for each agent, we combine the various component scores into a total, *TS* and, if the agent is performing poorly (in relation to the average score, *AvgS* in Line 7), we reconsider its boldness and vengefulness. Note that this average score is established through lines 1-5 in the algorithm.

Reinforcement learning algorithms typically involve an exploration parameter that balances exploration and exploitation of behaviours, ensuring that agents try every possible strategy that is available to them. To allow a degree of exploration in our algorithm (similar to mutation in the original model's evolutionary approach, to provide comparability) and to enable an agent to step out of the learning trend (captured by a learning parameter, or step, discussed below, δ), we adopt an *exploration rate*, γ, which regulates adoption of random strategies from the available strategies universe (Line 8). If an agent does not explore then, if defection is the cause of a low score (Line 12), the agent decreases its boldness, and increases it otherwise. Similarly, agents decrease their vengefulness if they find that the effect of punishing is worse than the effect of not punishing (Line 22), and increase vengefulness if the situation is reversed. As both *PS* and *POS* represent the result of two mutually exclusive actions, their difference for a particular agent determines the change to be applied to vengefulness. For example, if $PS > POS$, then punishment has some value, and vengefulness should be increased.

Finally, given a decision on whether to modify an agent's strategy, the degree of the change, or *learning step* (δ), must also be considered. Since vengefulness and boldness have eight possible values from $\frac{0}{7}$ to $\frac{7}{7}$, we adopt the conservative approach of increasing or decreasing by one level at each point, corresponding to a learning rate of $\delta = \frac{1}{7}$. Thus, an agent with boldness of $\frac{5}{7}$ and vengefulness of $\frac{3}{7}$ that decides to defect less and punish more will decrease its boldness to $\frac{4}{7}$ and increase its vengefulness to $\frac{4}{7}$.

Algorithm 4: learn(γ, δ)

```
$Temp = 0$
for each agent $i$ do
    $TS_i = TS_i + DS_i + PS_i + POS_i$
    $Temp = Temp + TS_i$
$AvgS = Temp/no\_agents$
for each agent $i$ do
    if $TS_i < AvgS$ then
        if explore($\gamma$) then
            $B_i = random()$
            $V_i = random()$
        else
            if $DS_i < 0$ then
                if $B_i - \delta < 0$ then
                    $B_i = 0$
                else
                    $B_i = B_i - \delta$
            else
                if $B_i + \delta > 1$ then
                    $B_i = 1$
                else
                    $B_i = B_i + \delta$
            if $PS_i < POS_i$ then
                if $V_i - \delta < 0$ then
                    $V_i = 0$
                else
                    $V_i = V_i - \delta$
            else
                if $V_i + \delta > 1$ then
                    $V_i = 1$
                else
                    $V_i = V_i + \delta$
```

5.3 Evaluation

The algorithm is designed to mimic the behaviour of Axelrod's evolutionary approach as much as possible, while relaxing Axelrod's unrealistic assumptions. This allows us to replicate Axelrod's results and investigate his approach in more realistic problem domains, as shown in Figure 16. The parameter set-up used in all experiments carried out in this section is shown in Table 4, with the addition of parameters related to the *BV Learning* algorithm.

The analysis of a sample run reveals that agents with low vengefulness and agents with high boldness start changing their strategies. Here, agents with high boldness defect frequently, and are punished as a result, leading to a very low *DS*, in turn causing these agents to decrease their boldness. Agents with low vengefulness do not punish and are consequently frequently metapunished. As a result, their *PS* is much better (lower in magnitude) than their *POS*, causing them to increase their vengefulness. The population eventually converges to comprise only agents with high vengefulness and low boldness. While noise is still introduced via the exploration rate causing random strategy adoption, the learning capability enables agents with such random strategies to adapt quickly to the trend of the population.

As before, we also consider the problem of ensuring that an original defection is observed in order to provide a metapunishment. Introducing this constraint

Table 4 Parameter initialisation

Term	Description	Value
i, j	Individuals	A numerical index to identify individual agents
S	Probability of a defection being seen by any given individual	Uniform distribution from 0 to 1
B_i	Boldness of i	Uniform distribution from $\frac{0}{7}$ to $\frac{7}{7}$
V_i	Vengefulness of i	Uniform distribution from $\frac{0}{7}$ to $\frac{7}{7}$
T	Player's temptation to defect	$+3$
H	Hurt suffered by others as a result of an agent's defection	-1
P	Cost of being punished	-9
E	Enforcement cost, i.e. cost of applying punishment	-2
$P\acute{}$	Cost of being punished for not punishing a defection	-9
$E\acute{}$	Cost of punishing someone for not punishing a defection	-2
δ	*learning step*	$\frac{1}{7}$
γ	*exploration rate*	0.01

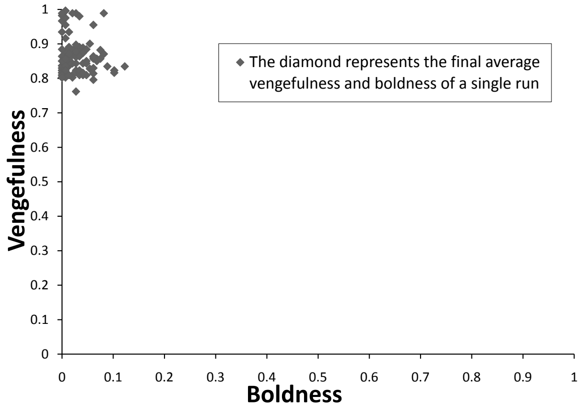


Fig. 16 Strategy improvement (with $\gamma = 0.01$), 1,000,000 timesteps

into our new algorithm, we ran experiments over different periods, with results indicating that norm establishment is robust in all runs. An example run for 1,000,000 timesteps is shown in Figure 17. This is because agents that use this new learning algorithm only change their strategy incrementally without wholesale change at any single point. The effect of a mutant with low vengefulness is not significant since, while the mutant might survive for a short period and cause some agents to change their vengefulness, any such change will be slight. It thus does not prevent such agents from detecting the mutant subsequently, in turn causing the mutant to increase its vengefulness.

Some work has already explored the use of learning in the context of distributed (non-centralised) punishment where agents are bold and vengeful at the same time [41]. This is interesting because it shows that punishment does not entail norm emergence (in the sense of stable collective behaviour coherent with punishments)

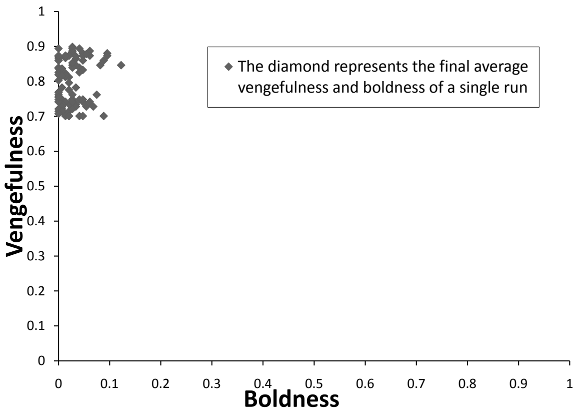


Fig. 17 Strategy improvement with defection observation constraint (with $\gamma = 0.01$), 1,000,000 timesteps

even when it is strongly applied. In addition, together with later work that also considers vicarious reinforcement [40], by which agents observe the behaviour of others in response to rewards, for example, we can see a different perspective on some of the same techniques. This vicarious learning of new strategies by which agents adopt the strategy of better performing agents has been studied (for the legal domain) where, if combined with reinforcement learning, it tends to increase the speed of convergence so that agents quickly adopt the most useful behaviours. Moreover it is not clear that this necessarily supports norm establishment, but this is a combination of techniques that we have not explored since, in our model it has not proven necessary to combine them. Nevertheless, this could provide a useful avenue for future investigation. More generally, while interesting and relevant to the larger questions surrounding learning in the context of norm emergence, the underlying model and assumptions in these works are rather different to our own, in which we have sought to retain some degree of faithfulness to Axelrod's original intent, with incremental deviations to better ensure comparability.

6 Conclusion

Axelrod's original work suggested that without metanorms, whether or not laws emerge depends on the particular social composition of the initial group of agents in terms of their boldness and vengefulness. His introduction of metanorms was aimed at addressing this, enabling societies to guard against normative collapse and seeing norms emerge from the interactions of the individuals within those societies. In the context of his model, in which these individuals are characterised by vengefulness and boldness, this brings about a situation tending to low boldness and high vengefulness, but with scenarios in which there is susceptibility to the arrival of a few bold individuals that can bring about norm collapse. While real societies with metanorms of this sort are rare, compliance (and playing a role in ensuring compliance) is often seen as a duty. In a low vengefulness society citizens prefer to discharge this duty by paying someone else to enforce the laws: this is cheaper and easier and more congenial, and it spreads the risk of incurring costs of punishing. In real societies, we thus see the emergence of a social contract with a central enforcement paid for from a collective levy (initially subscription for this specific purpose, but as the functions of the state increase, taxation). This leads to the development of a central authority and *the rule of law*, but this is difficult to bring about in computational societies.

Thus, in computational systems of self-interested autonomous agents we often need to establish cooperative norms to ensure the desired functionality without this centralisation. Axelrod's work on norm emergence gives valuable insight into the mechanisms and conditions in which such norms may be established. However, there are two major limitations. First, as we have shown, and explained in detail, norms collapse even in the metanorms game for extended runs. Second, the model suffers from limitations relating to assumptions of omniscience. In response, this paper has: (i) provided an in-depth analysis of Axelrod's results, revealing the effects of mutation and reproduction; (ii) investigated those aspects of Axelrod's model that are unreasonable in real-world domains; and (iii) proposed *BV learning* as an alternative mechanism for norm establishment that avoids these limitations.

More specifically, we replaced the evolutionary approach with a learning interpretation in which, rather than remove and replicate agents, we allow them to learn from others. Two techniques were considered: copying from a single agent and copying from a group. The former suffers the same problems of long term norm collapse associated with Axelrod's approach but, by avoiding strategies that only perform well in restricted settings, the latter addresses the problems and brings about norm establishment. In addition, we addressed Axelrod's assumption of omniscience, in which agents considering metapunishment are not explicitly required to *see* the original defection. By doing so, however, the metapunishment activity in the population, for stabilising an established norm, decreases and leads to norm collapse.

Since learning strategies from *others* (either individuals or groups) is unable to establish norms for cooperation (and is, in addition, unrealistic since it assumes that agent strategies are not private), we have developed an alternative, *BV learning*, in which agents learn from their *own* experiences. Through this approach we have shown that not only it is possible to avoid the unrealistic assumption of knowledge of others' strategies, but also that by enabling individuals to incrementally change their strategies we can avoid norm collapse, even with observation constraints on metapunishment.

Of course, this model remains a simple one that is still constrained to limited underlying structures. Our aim is to focus on applying the model to interaction networks in order to analyse how different network structures can impact on the achievement of norm emergence. In particular, our current model is limited in that the algorithm relies on agents comparing their own score to the average score of all other agents to determine if learning is warranted. This constrains our ability to move towards making Axelrod's model more suitable for real-world distributed systems and, in consequence, future work aims to enable agents to estimate their learning needs based on their own, individual, experience by monitoring their past performance. Moreover, we also seek to investigate the possibility of integrating *dynamic* punishments, rather than the current static ones (that are fixed regardless of what has happened), by which agents can modify the punishments they impose on others according to available information about the severity of violation, or according to whether the violating agent is a repeat offender, and if so, how many times.

There is also another interesting characteristic of this work. The scenarios discussed in the paper consider some actions that can be seen as stochastically dependent. In this respect, defecting is largely dependent on the probability of being seen, and the lower this probability, the greater the possibility that an agent will defect if it has sufficient boldness. Similarly, the actions of punishment and metapunishment are stochastically dependent on the probability of being seen, whereby an agent A can only punish or metapunish another agent B if it observes the defection of B, and it is sufficiently vengeful. An interesting avenue to explore, therefore, could be to provide a more detailed and considered analysis of the correlations here, but this is outside the scope of the current work in this paper.

Acknowledgements

We are immensely grateful to the anonymous reviewers, whose detailed and insightful comments have improved the content and presentation of this paper. Without their valuable work, this would be a much poorer paper, and its value much diminished.

References

1. R. Axelrod. *The evolution of cooperation.* Basic Books, 1984.
2. R. Axelrod. An evolutionary approach to norms. *The American Political Science Review*, 80(4):1095–1111, 1986.
3. T. J. M. Bench-Capon. Analysing norms with transition systems. In *JURIX 2014: Proceedings of the Twenty-Seventh Annual Conference on Legal Knowledge and Information Systems*, pages 29–38. IOS Press, 2014.
4. Cristina Bicchieri. *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge University Press, 2005.
5. K. Binmore. Review of Complexity and Cooperation by Robert Axelrod. *Journal of Artificial Societies and Social Situations*, 1(1):82, 1998.
6. K. Binmore, J. Gale, and L. Samuleson. Learning to be imperfect: The ultimatum game. *Games and Economic Behavior*, 8:56–90, 1995.
7. G. Boella, L. van der Torre, and H. Verhagen. Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2-3):71–79, 2006.
8. M. Boman. Norms as constraints on real-time autonomous agent action. In *Multi-Agent Rationality: Proceedings of the 8th European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, volume 1237 of *Lecture Notes in Computer Science*, pages 36–44. Springer, 1997.
9. E. Borenstein and E. Ruppin. Enhancing autonomous agents evolution with learning by imitation. *Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behavior*, 1(4):335–348, 2003.
10. M. Bowling and M. Veloso. Rational and convergent learning in stochastic games. In *IJCAI 2001: Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 1021–1026, 2001.
11. C. Castelfranchi and R. Conte. From conventions to prescriptions. towards a unified theory of norms. *Artificial Intelligence and Law*, 7:323–340, 1999.
12. R. Conte and C. Castelfranchi. Understanding the functions of norms in social groups through simulation. In N. Gilbert and R. Conte, editors, *Artificial Societies The Computer Simulation of Social Life*, pages 252–267. UCL Press, 1995.

13. K. Dautenhahn and C. L. Nehaniv, editors. *Imitation in animals and artifacts.* MIT Press, Cambridge, MA, USA, 2002.
14. J. Delgado. Emergence of social conventions in complex networks. *Artificial Intelligence*, 141(1-2):171–185, October 2002.
15. J. Delgado, J. M. Pujol, and R. Sangesa. Emergence of coordination in scale-free networks. *Web Intelligence and Agent Systems*, 1:131–138, 2003.
16. F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999.
17. B. Druzin. Law without the state: The theory of high engagement and the emergence of spontaneous legal order within commercial systems. *Georgetown Journal of International Law*, 41(3):559–620, 2010.
18. E. Durkheim. *De la division du travail social.* Les classiques des sciences sociales. Université due Québec à Chicoutimi, 1893.
19. E. Ehrlich. *Fundamental Principles of the Sociology of Law.* Transaction Publishers, 1913.
20. J. M. Epstein. Learning to be thoughtless: Social norms and individual computation. *Computational Economics*, 18(1):9–24, 2001.
21. H. Franks, N. Griffiths, and A. Jhumka. Manipulating convention emergence using influencer agents. *Autonomous Agents and Multi-Agent Systems*, pages 1–39, 2012.
22. J. M. Galan and L. R. Izquierdo. Appearances can be deceiving: Lessons learned reimplementing Axelrod's evolutionary approach to norms. *Journal of Artificial Societies and Social Simulation*, 8(3), 2005.
23. J.P. Gibbs. Norms: The problem of definition and classification. *The American Journal of Sociology*, 70(5):586–594, 1965.
24. R. Guerraoui, K. Huguenin, A. Kermarrec, and M. Monod. On Tracking Freeriders in Gossip Protocols. In *P2P 2009: Proceedings of the 9th International Conference on Peer-to-Peer Computing*, 2009.
25. G. Hayes and J. Demiris. A robot controller using learning by imitation. In *Proceedings of the 2nd International Symposium on Intelligent Robotic Systems*, pages 198–204, 1994.
26. J. M. Helmhout, H. W. M. Gazendam, and R. J. Jorna. Control over emergence. In *AISB 2008: Proceedings of the Convention: Communication, Interaction, and Social Intelligence*, pages 1–8, 2008.
27. M. Kandori, G. J. Mailath, and R. Rob. Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1):29–56, 1993.
28. J. Kittock. Emergent conventions and the structure of multi–agent systems. In *Lectures in Complex systems: the proceedings of the 1993 Complex systems summer school, Santa Fe Institute Studies in the Sciences of Complexity Lecture Volume VI, Santa Fe Institute*, pages 507–521. Addison-Wesley, 1995.
29. R. Krishnan, D. M. Smith, Z. Tang, and R. Telang. The impact of free-riding on peer-to-peer networks. In *HICSS '04: Proceedings of the 37th Annual Hawaii International Conference on System Sciences.* IEEE Computer Society, 2004.
30. K. Kulakowski and P. Gawronski. To cooperate or to defect? altruism and reputation. *Physica A: Statistical Mechanics and its Applications*, 388(17):3581–3584, 2009.
31. K. Lakkaraju and L. Gasser. Norm emergence in complex ambiguous situations. In *COIN 2008: Proceedings of the AAAI Workshop on Coordination, Organizations, Institutions, and Norms*, 2008.
32. M. Lloyd-Kelly, K. Atkinson, and T. J. M. Bench-Capon. Fostering co-operative behaviour through social intervention. In *SIMULTECH 2014: Proceedings of the 4th International Conference On Simulation And Modeling Methodologies, Technologies And Applications*, pages 578–585. IEEE, 2014.
33. F. López y López and M. Luck. Modelling norms for autonomous agents. In E. Chávez, J. Favela, M. Mejía, and A. Oliart, editors, *Proceedings of The 4th Mexican Conference on Computer Science*, pages 238–245. IEEE Computer Society, 2003.
34. P. Mukherjee, S. Sen, and S. Airiau. Emergence of norms with biased interactions in heterogeneous agent societies. In *Web Intelligence and Intelligent Agent Technology Workshops, 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 512–515, 2007.
35. M. Nakamaru and U. Dieckmann. Runaway selection for cooperation and strict-and-severe punishment. *Journal of theoretical biology*, 257(1):1–8, 2009.
36. M. Neumann. The cognitive legacy of norm simulation. *Artificial Intelligence and Law*, 20:339–357, 2012.

37. M. J. Prietula and D. Conway. The evolution of metanorms: quis custodiet ipsos custodes? *Computational and Mathematical Organization Theory*, 15(3):147–168, 2009.
38. M. Rheinstein. *Max Weber on Law and Economy in Society.* Harvard University Press, 1954.
39. R. Riolo, M. Cohen, and R. Axelrod. Evolution of cooperation without reciprocity. *Nature*, 414:441–443, 2001.
40. R. Riveret, G. Contissa, D. Busquets, A. Rotolo, J. Pitt, and G. Sartor. Vicarious reinforcement and ex ante law enforcement: a study in norm-governed learning agents. In *International Conference on Artificial Intelligence and Law, ICAIL '13, Rome, Italy, June 10-14, 2013*, pages 222–226, 2013.
41. R. Riveret, A. Rotolo, and G. Sartor. Probabilistic rule-based argumentation for norm-governed learning agents. *Artificial Intelligence and Law*, 20(4):383–420, 2012.
42. N. Salazar, J. A. Rodriguez-Aguilar, and J. L. Arcos. Robust coordination in large convention spaces. *AI Communications*, 23:357–372, December 2010.
43. B. T. R. Savarimuthu, S. Cranefield, M. Purvis, and M. Purvis. Role model based mechanism for norm emergence in artificial agent societies. In *COIN '07: Proceedings of the International Workshop on Coordination, Organization, Institutions and Norms*, pages 1–12, 2007.
44. Bastin Tony Roy Savarimuthu and Stephen Cranefield. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent Grid Systems*, 7(1):21–54, 2011.
45. O. Sen and S. Sen. Effects of social network topology and options on norm emergence. In J. Padget, A. Artikis, W. Vasconcelos, K. Stathis, V. da Silva, E. Matson, and A. Polleres, editors, *Coordination, Organizations, Institutions and Norms in Agent Systems V*, volume 6069 of *Lecture Notes in Computer Science*, pages 211–222. Springer Berlin / Heidelberg, 2010.
46. S. Sen and S. Airiau. Emergence of norms through social learning. In *IJCAI 2007: Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1507–1512. Morgan Kaufmann, 2007.
47. Y. Shoham and M. Tennenholtz. Emergent Conventions in Multi-Agent Systems: Initial Experimental Results and Observations (Preliminary Report). In *Proceedings of the 3rd International Conference on KR&R*, pages 225–232, 1992.
48. Y. Shoham and M. Tennenholtz. Co-learning and the evolution of social acitivity. Technical report, Stanford, CA, USA, 1994.
49. Y. Shoham and M. Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94:139–166, 1997.
50. T. Slembeck. Reputations and fairness in bargaining - experimental evidence from a repeated ultimatum game with fixed opponents. Experimental, EconWPA, May 1999.
51. M. Song, R. Chen, and j. An. Social conventions to promise learning convergence. In *FSKD '07: Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 660–662, Washington, DC, USA, 2007. IEEE Computer Society.
52. E. Ullman-Margalit. *The Emergence of Norms.* Clarendon Press, Oxford, 1977.
53. P. Urbano, J. Balsa, L. Antunes, and L. Moniz. Force versus majority: A comparison in convention emergence efficiency. In *Coordination, Organizations, Institutions and Norms in Agent Systems IV: COIN 2008 International Workshops, COIN@AAMAS 2008, Estoril, Portugal, May 12, 2008. COIN@AAAI 2008, Chicago, USA, July 14, 2008. Revised Selected Papers*, pages 48–63, 2008.
54. D. Villatoro, G. Andrighetto, J. Sabater-Mir, and R. Conte. Dynamic sanctioning for robust and cost-efficient norm compliance. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 414–419, Barcelona, 2011. AAAI Press.
55. D. Villatoro, J. Sabater-Mir, and S. Sen. Social instruments for robust convention emergence. In Toby Walsh, editor, *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 420–425. AAAI Press, 2011.
56. D. Villatoro, S. Sen, and J. Sabater-Mir. Topology and memory effect on convention emergence. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technologies*, pages 233–240. IEEE, 2009.
57. A. Walker and M. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In V. Lesser, editor, *Proceedings of the First International Joint Conference on Multi Agent Systems)*, pages 384–389, 1995.
58. C.J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

59. M. Yang, Z. Zhang, X. Li, and Y. Dai. An empirical study of free-riding behavior in the maze P2P file-sharing system. In *IPTPS '05: Proceedings of the 4th International Workshop on Peer-to-Peer Systems*, pages 182–192. Springer, 2005.