

An Analysis of Norm Emergence in Axelrod’s Model

Samhar Mahmoud¹, Nathan Griffiths², Jeroen Keppens¹, Michael Luck¹

¹ Department of Computer Science, King’s College London, London WC2R 2LS, UK. email: samhar.mahmoud@kcl.ac.uk

² Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK.

Abstract. Norms are a valuable mechanism for establishing coherent cooperative behaviour in decentralised systems in which no central authority exists. One of the most influential formulations of norm emergence was proposed by Axelrod [1], who defined a model of norms and metanorms that enables norm establishment in populations of self-interested individuals. 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. First, we explore the dynamics of norm emergence and the occurrence of norm collapse when applying the model over extended durations. Second, we investigate in detail the reasons for norm collapse in extended applications of the model and show that both the level of mutation in the population and the precise nature of the reproduction mechanism are significant. 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.

1 Introduction

As has been suggested by many (e.g., [2–4, 8, 12]) *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 (e.g., [5, 6, 10, 13, 11]) perhaps most notably Axelrod, whose seminal paper in 1986 offered a model of norms and metanorms [1] that has since been further investigated [7, 9].

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. An agent’s behaviour is assessed by means of a careful system of scoring that simulates the potential rewards and penalties associated with norm violation and enforcement. The agent population is evolved 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 determined by 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. Key to Axelrod's model is the notion of *metanorms*, secondary norms that help to enforce compliance with primary norms by punishing agents that do not themselves punish a defector. By using metanorms, Axelrod was able to establish norms in his experiments.

However, as was more recently shown by Galan and Izquierdo [7], Axelrod's 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, showing that their results, too, rely on some assumptions and conditions, and we provide a further analysis of Axelrod's model, drawing out some important considerations for the establishment of norms more generally.

The paper begins with a description of Axelrod's *Norms Game*, followed by a more detailed analysis of the results than provided elsewhere. Then, in Section 4, the duration of the game, a critical part of Galan and Izquierdo's analysis, is reviewed, with different results, leading to a new consideration of the circumstances for norm collapse (when norms are not established). In Section 4, the impact of the reproduction policy is analysed, and in Section 5, a consideration of the impact of mutation is provided. The paper concludes with a discussion and conclusions on the significance of these results.

2 Axelrod's Model

2.1 The Norms Game

Axelrod's *norms game* adopts an evolutionary approach in which successful strategies multiply over generations, potentially leading to convergence on norms. 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); and 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. If $S < B_i$, then 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 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 8 possible values, they need three bits to be represented, to which mutation is applied (by flipping a bit) whenever an agent is reproduced, with a 1% chance.

Axelrod's experiment was run five times, each with 100 generations. Two runs resulted in high average boldness and almost zero average vengefulness so there was no norm emergence at all, two other runs gave low boldness and vengefulness, but only the final run had a high level of vengefulness and very low boldness, indicating a partial establishment of a norm against defection. In response, Axelrod considered an additional mechanism to support norms in his *metanorm* model.

2.2 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 defection itself, and others in turn may observe this defection (with probability S) and apply a punishment to the non-enforcing agent. 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 $E' = -2$. 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.

3 Analysis of Axelrod's Model

In seeking to replicate Axelrod's results [1], it becomes clear that some assumptions need to be made, about which Axelrod says nothing. First, the model does not specify how the constant population level is maintained after reproduction, when there are three possible scenarios. (i) The new population is smaller than the original. In our reimplementation, additional agents are randomly selected from the resulting population and replicated. (ii) The new population is equal in size to the original, in which case nothing new is needed. (iii) The new population is larger than the original. In this case, our position is to determine the required number of agents at random from the relevant set for reproduction. Second, we assume the score of each agent is set to 0 at the beginning of each generation.

We repeated Axelrod's experiments, running the norms game 10 times, with results as shown in Figure 1, where the diamonds represent the value of the mean average boldness and vengefulness of the final generation population. This is similar to Axelrod, with one run having high vengefulness and low boldness, two runs with exactly the

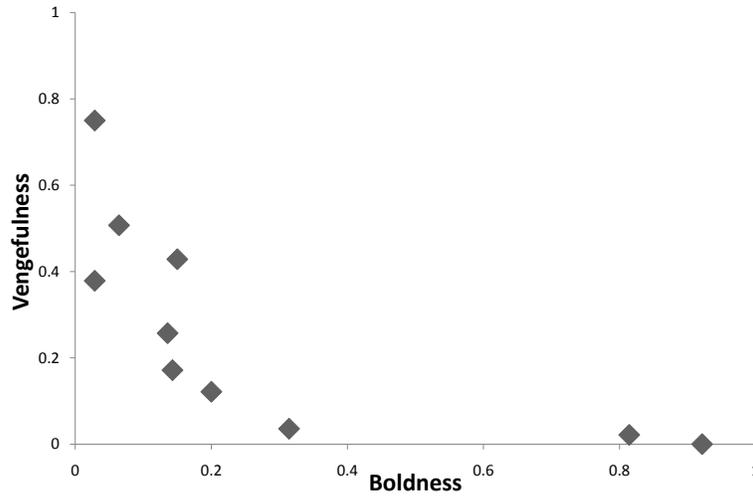
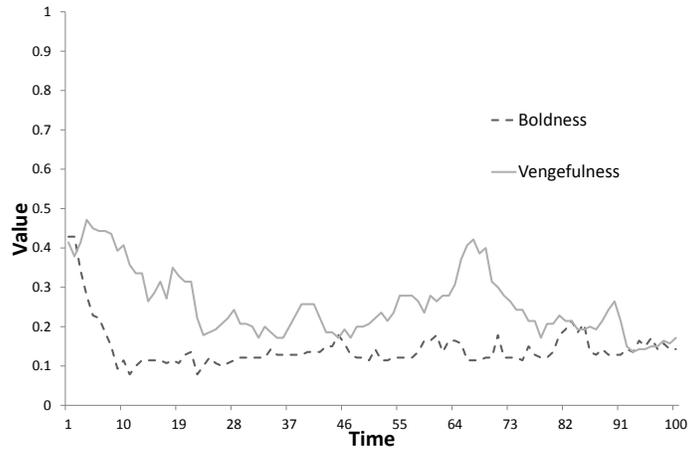


Fig. 1: Norms game overall results

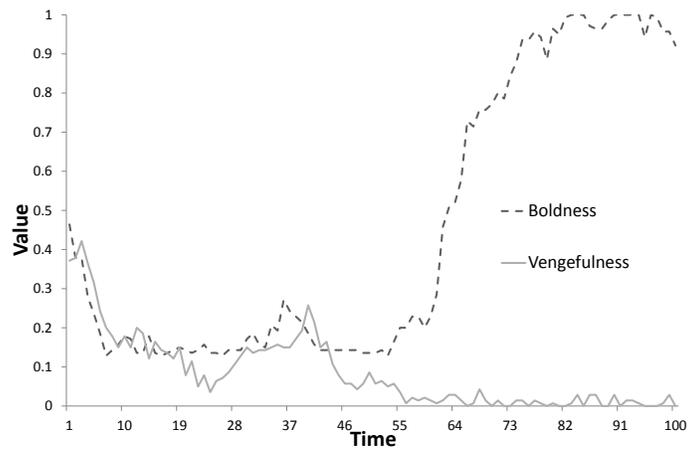
opposite (high boldness and low vengefulness), and all other runs with low values for both boldness and vengefulness.

In order to establish how these results arise, changes to boldness and vengefulness for each individual were monitored; Figure 2 provides some sample graphs illustrating this, showing the average boldness and vengefulness as they 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 $\{\frac{0}{7}, \dots, \frac{7}{7}\}$). In the early stages, boldness decreases slightly, indicating that individuals with higher boldness are eliminated. This is because high boldness causes an agent to defect, yet defecting 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. Finally, the values stabilise at particular low values for both boldness and vengefulness until the end of the run.

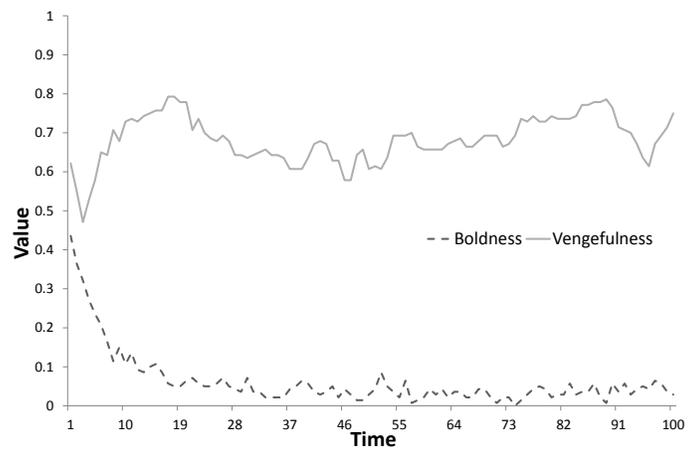
In the cases that result in high boldness and low vengefulness (an example run is shown in Figure 2(b)), the run starts as before, with both values reducing. However, around the 60th generation, the value of boldness increases sharply until it reaches 1, where it remains. This can be explained by a dramatic 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: when the high boldness phase ends, only individuals with high vengefulness remain, so there are no individuals with low boldness



(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 runs

and low vengefulness, and those with high vengefulness and low boldness survive and dominate.

By introducing metanorms, Axelrod aimed to address the problems identified above. Our own simulation (see Figure 3) provides similar results to Axelrod but again, a deeper analysis is required. As shown in Figure 4, in this new game, the population starts eliminating high boldness individuals as before, but now also eliminates low vengefulness individuals. The latter trend results from the metanorm game in 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.

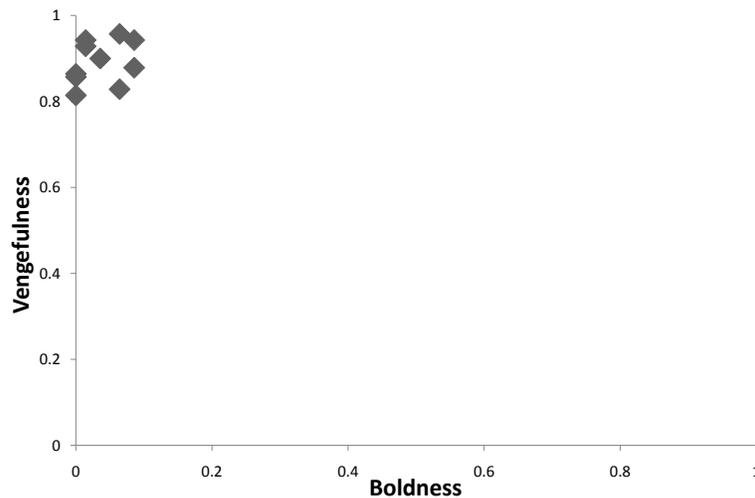


Fig. 3: Metanorms game overall results

4 Game Duration

In analysing the model, we repeated these experiments over a *long duration*, 1,000,000 generations (and 10 runs), as opposed to 100 generations, to provide a stronger analysis. In our norm 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 opportunities for this to occur and, as shown in Figure 5, this always leads to norm collapse.

In undertaking their own analysis of Axelrod's metanorm model, Galan and Izquierdo [7] 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

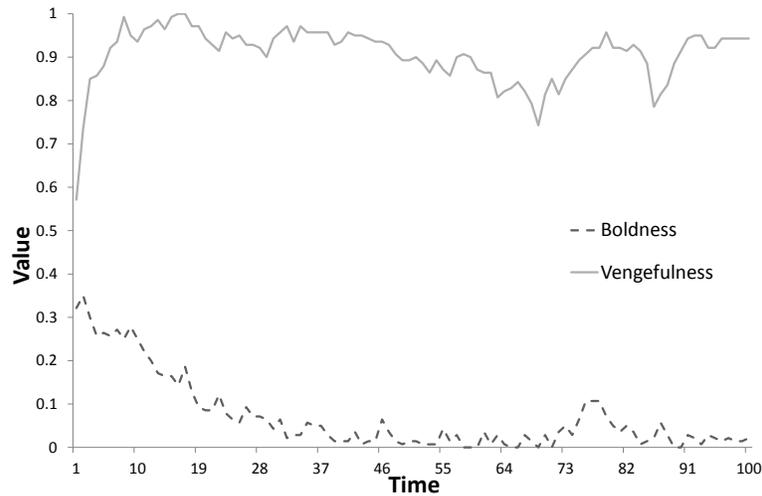


Fig. 4: Metanorms game analysis

to Axelrod's *norm establishment*. According to Galan, 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. Galan's results suggest that metanorms are not as useful as it might seem from Axelrod's results.

In analysing these cases to determine the reasons for this result, it's 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 *A* 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 *B*, with average vengefulness remaining high, *B* 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 *A* may not punish *B* 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, *A* will not be punished for non-punishment, but in the latter case, *A* might be if it is seen by others. However, we know that the probability of being seen is low because agent *B* has defected (and $S < B$ for defection to take place). In this case, *B* is eliminated, while *A* remains, because the likelihood of *B*'s defection being seen by just one agent is relatively high, but the likelihood of agent *A*'s non-punishment being seen requires first *B*'s defection being seen by *A*, and then *A*'s non-punishment being seen by others, the combination of these being extremely unlikely.

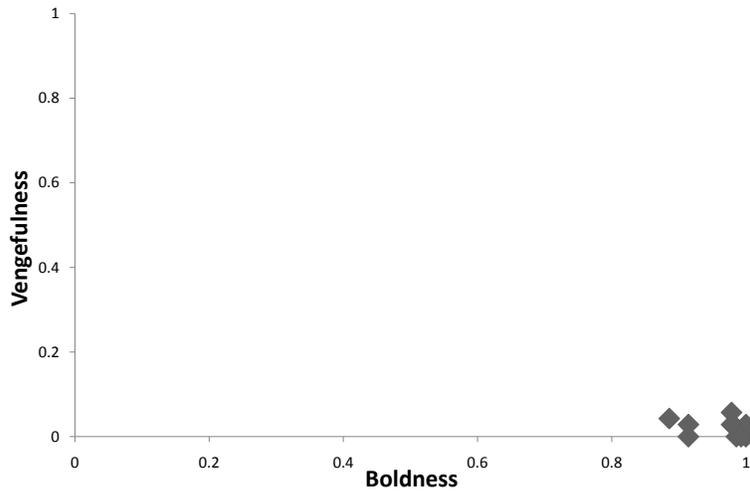


Fig. 5: Norms game for 1,000,000 generations

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.

This analysis suggests that both mutation and sanctioning 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 model which, in certain circumstances, and with very subtle change, can give a very different outcome. We consider this next.

5 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. At this point, the population stabilises at the same level of vengefulness and boldness until the end of the run. Now, when all individuals have boldness around 0, defection rarely happens, and their scores (which change only when agents are hurt, punished or enforce) 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.

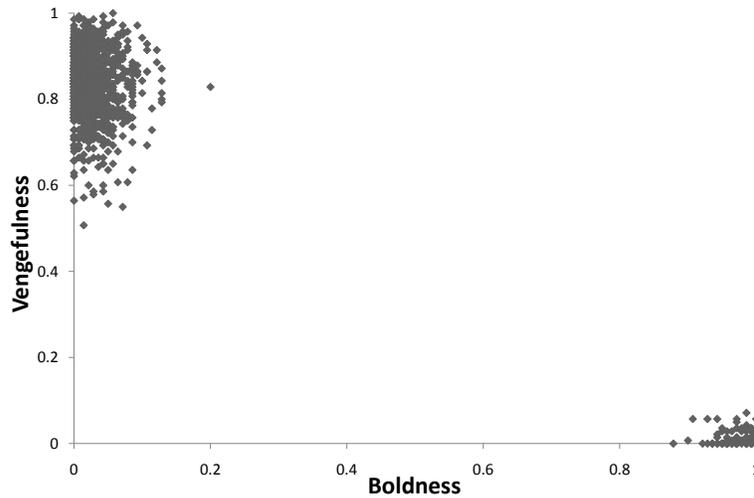


Fig. 6: Metanorms game for 1,000,000 generations

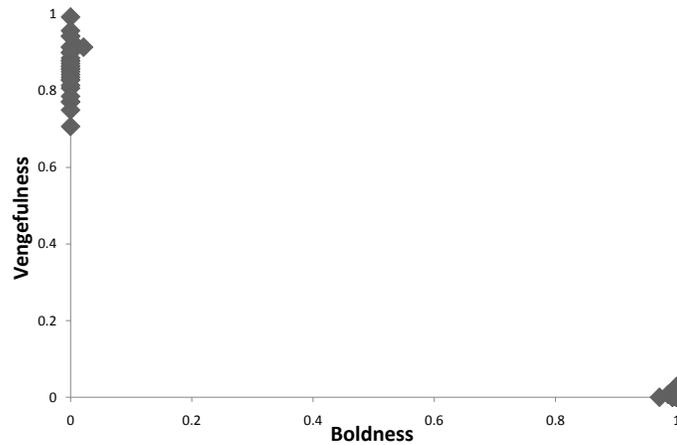
This means that all agents are replicated and then the new generation is selected at random from this pool of agents, as described in Section 3.1

It is this key point that is critical to these striking results. According to Axelrod's rules, agents in this situation should be replicated *twice* when forming the new generation, because all agents have a score equal to the average plus one standard deviation. However, duplicating the individuals in this case does not seem sensible since it does not fulfil the original purpose suggested by Axelrod, of giving individuals with better scores a greater chance of survival. If, in contrast, as we have done in our own simulation for this special case alone, we only replicate once, then the results of the metanorm game with many more generations are similar to those of Galan, 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.

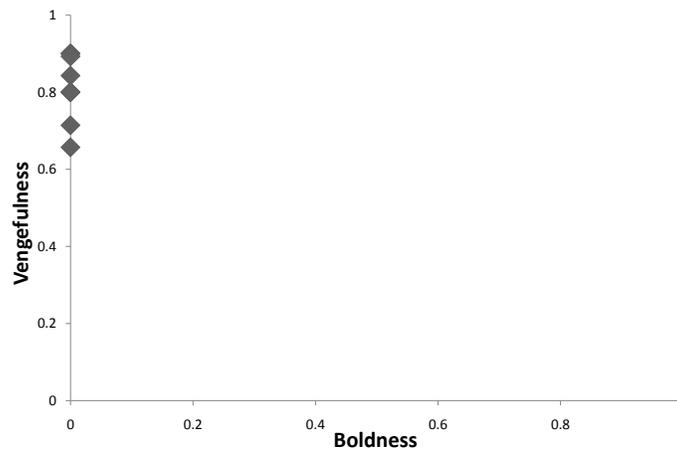
Replication of an entire population of non-defecting agents increases the likelihood of significant fluctuations in vengefulness over subsequent generations. For example, in one phase of a run using Galan's approach in which all agents have 0 boldness, five agents have vengefulness of 0, eleven with 1 and four with 0.8, the next generation includes eight agents with vengefulness of 0, seven agents with 1, and five 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, replication again makes this worse. However, replication could cause the opposite, increasing the number 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. Until vengefulness levels fluctuate back in an upwards situ-

ation, 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 duplicated in subsequent iterations of the game. The end result of the



(a) 400,000 generations and 0.001 mutation rate



(b) 400,000 generations and 0.0001 mutation rate

Fig. 7: Metanorms game with different mutation rates

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.

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

6 Mutation

As discussed above, mutation is significant in determining when norm collapse occurs. Galan and Izquierdo [7] 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 toward norm collapse in about 25,000 generations as opposed to 300,000 with 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 50 runs of 400,000 generations, using the mutation rate of 0.001 suggested by Galan, with results shown in Figure 7(a). Clearly, these results support Galan's argument: 20 of 50 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 runs ending in norm establishment, shown in Figure 7(b).

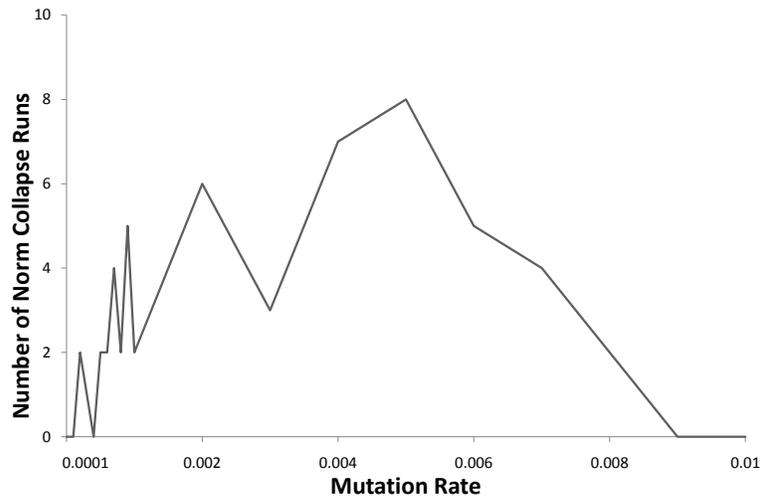


Fig. 8: Metanorms game: 1m generations; 0.0001–0.01 mutation rate

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. 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. Figure 9 illustrates the result of a no mutation norms game that consists of 40 runs, with 1,000,000 generations each. As expected, removing mutation avoids norm collapse and leaves the population with the other two situations. In the metanorms game,

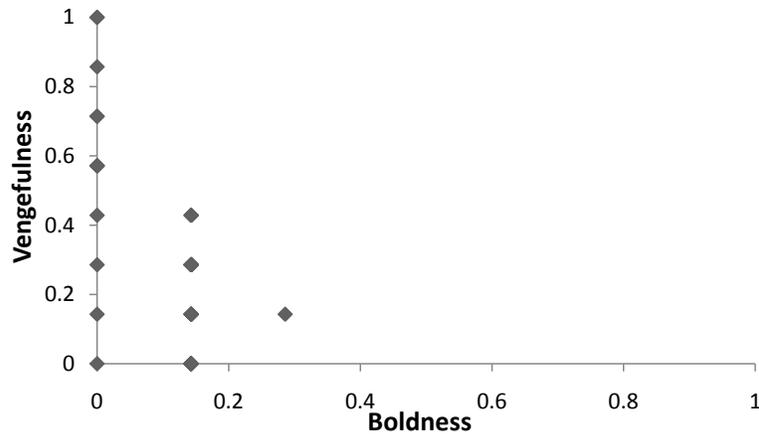


Fig. 9: No mutation norms game: 1,000,000 generations

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 100 runs and Figure 10(b) for 1,000,000 generations and 50 runs. Surprisingly, the result was 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: 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 longer has no impact and there is no change to the population once the levels of vengefulness and boldness stabilise.

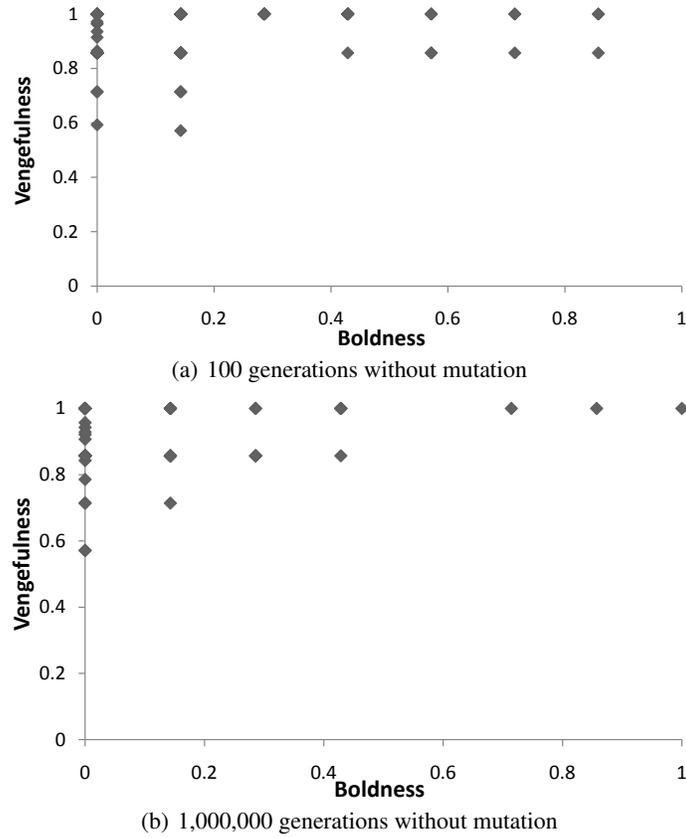


Fig. 10: Metanorms game with different mutation rates

7 Discussion and Conclusions

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, and 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 through the paper, 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 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's 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, just as the run in Figure 3.

While we have analysed the impact of duration, reproduction policy and mutation rate on Axelrod's game, some of the results, such as the exact relation of mutation rate to norm collapse also require further analysis than is possible in this paper. In addition, we have only addressed some of the potentially relevant issues both due to space constraints and time in undertaking our analysis. For example, some results in

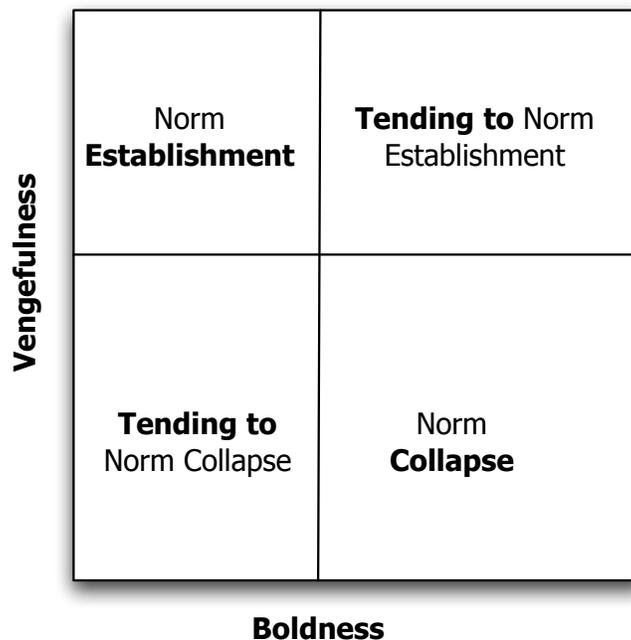


Fig. 11: Characterising the vengefulness-boldness space

the paper suggest that the sanctioning structure does not allow a population to recover once it stabilises on high boldness and low vengefulness. In such situations, agents with average vengefulness score lower than others because of the high enforcement costs, which might be addressed by introducing a larger gap between enforcement and punishment costs. While Galan and Izquierdo [7] experiment with a set of reduced values for *meta* punishment and *meta* enforcement (but preserving the ratio), leading to norm collapse much earlier, their analysis is not extensive. Indeed, by preserving the

ratio of costs but at lower values, it seems obvious that norms will collapse, because metanorms remain ineffective, as agents may pay a much higher price for enforcement than the punishment value they receive for not punishing.

Clearly, further work on the impact of sanctions on the model may be valuable in determining the exact relationship to norm collapse and establishment. Other relevant issues may include population size, temptation value (as suggested by Galan), and so on. All these are currently the subject of further analysis as our own work progresses.

References

1. R. Axelrod. An evolutionary approach to norms. *The American Political Science Review*, 80(4):1095–1111, 1986.
2. G. Boella, L. van der Torre, and H. Verhagen. Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2-3):71–79, 2006.
3. M. Boman. Norms as constraints on real-time autonomous agent action. In *Multi-Agent Rationality: Proceedings of the Eighth European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, volume 1237 of *Lecture Notes in Computer Science*, pages 36–44. Springer, 1997.
4. F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999.
5. J. M. Epstein. Learning to be thoughtless: Social norms and individual computation. *Computational Economics*, 18(1):9–24, 2001.
6. F. Flentge, D. Polani, and T. Uthmann. Modelling the emergence of possession norms using memes. *Journal of Artificial Societies and Social Simulation*, 4(4), 2001.
7. J. M. Galan and L. R. Izquierdo. Appearances can be deceiving: Lessons learned re-implementing Axelrod’s evolutionary approach to norms. *Journal of Artificial Societies and Social Simulation*, 8(3), 2005.
8. J.P. Gibbs. Norms: The problem of definition and classification. *The American Journal of Sociology*, 70(5):586–594, 1965.
9. M. J. Prietula and D. Conway. The evolution of metanorms: quis custodiet ipsos custodes? *Computational and Mathematical Organization Theory*, 15(3):147–168, 2009.
10. B. T. R. Savarimuthu, M. Purvis, M. Purvis, and S. Cranefield. Social norm emergence in virtual agent societies. In M. Baldoni, T. C. Son, M. B. van Riemsdijk, and M. Winikoff, editors, *Declarative Agent Languages and Technologies VI*, volume 5397 of *Lecture Notes in Computer Science*, pages 18–28. Springer, 2009.
11. 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.
12. 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 Fourth Mexican Conference on Computer Science*, pages 238–245. IEEE Computer Society, 2003.
13. T. Yamashita, K. Izumi, and K. Kurumatani. An investigation into the use of group dynamics for solving social dilemmas. In P. Davidsson, B. Logan, and K. Takadama, editors, *Multi-Agent and Multi-Agent-Based Simulation*, volume 3415 of *Lecture Notes in Artificial Intelligence*, pages 185–194. Springer, 2005.