Rerum Causae
Student journal of Philosophy

Leon Assaad (2021) *The Structural Evolution of Cooperation: Can Evolutionary Game Theory Teach Us About Morality?*, Rerum Causae, 12(1)

The Structural Evolution of Cooperation

Can Evolutionary Game Theory Teach Us About Morality?

Leon Assaad[1]

[1] University of Bayreuth

Abstract

This paper is a discussion of Jason McKenzie Alexander's book *The Structural Evolution of Morality*. In it he uses a set of evolutionary game theoretic models to expose a link between morality and rationality: The moral norms we uphold are successful heuristics to produce the best outcome for each member within a society.

I discuss and evaluate one of these models in detail: the dynamic network model, which suggests that dynamic social structure promotes cooperation in a Hobbesian state of nature. Even though interactions are represented by the Prisoner's Dilemma, the model's population of self-interested agents can evolve into universal cooperation. This indeed suggests that our real cooperative norms are in place because they are evolutionarily advantageous. However, does the model allow for inferences to true conclusions about real cultural evolution? I draw on Uskali M¨aki's evaluative framework and D'Arms' et al.'s guidelines for assessing the model's explanatory power. In conclusion: Although more sophisticated than many models of its kind, the dynamic network model does not allow for model-to-world inferences. It is no credible surrogate system for its target and can therefore not shed light on our cooperative norms.

1 Introduction

In *The Structural Evolution of Morality*, Jason McKenzie Alexander explores our moral norms. He uses evolutionary game theory as a tool to expose a link between morality and rationality: The moral norms we uphold serve as heuristics to produce the best outcome for each member within a society. Using increasingly more complex models of cultural evolution, Alexander puts our fundamental norms to the test. He concludes: Social structure favors the evolution of the moral principles we value. This indeed suggests that our moral norms not only indicate what we take to be the "right thing to do", they are also successful heuristics proven to be evolutionarily advantageous. Alexander uses a dynamic network model to show that social structure may prevent society's demise into a war of all against all, when placed into a Hobbesian state of nature. He claims this to demonstrate the aforementioned relation between rationality and morality for our norm of cooperation: When the situation is grim, not only do we value cooperation. The model shows that cooperation is also an evolutionarily advantageous strategy.

I evaluate this argument. I first explain how Alexander uses evolutionary game theory as a tool for inquiries into morality. I sketch the dynamic network model, which Alexander borrows from fellow game theorists Skyrms and Pemantle. In order for the model to explain anything about our real moral norms, it must be utilized to make a claim. I work out what exactly Alexander is aiming for and what explanatory power the model has for his purposes. Lastly, I discuss the model's effectiveness. I draw on Uskali Ma˙ki's evaluative framework and D'Arms' et al.'s guidelines for assessing Alexander's explanation. In conclusion: Even though Alexander's dynamic network is superior to other models of the kind, it fails to demonstrate that our moral norms are in fact successful heuristics in evolutionary settings.

2 Evolutionary Game Theory and Ethics

Because of its broad applicability game theory has become a tool for inquiries into morality. Two features make evolutionary game theory (EGT) an especially fitting tool: It models agents more realistically than standard game theory's hyperrational agents (Alexander 2019) and it studies their behavior in repeated interactions (Verbeek and Morris 2018).

2.1 Bounded Rationality

Standard game theory is a subpar description of real human behavior: Con- trary to the discipline's assumption, we fall short of the preferential axioms assumed for the homo economicus.[1] For a realistic conception of agency we cannot use the traditional theory of expected utility and must therefore abandon standard game theory. Alexander proposes a conception of bounded rationality: Agents rely on "fast and frugal" heuristics when tackling decision problems (Alexander 2007, p. 07). They are satisficing agents: Each individual has an aspiration level he wishes to attain. Lastly, boundedly rational agents maximize their preference satisfaction, rather than their utility. This more realistic model of individual agency calls for an equally realistic model of social dynamics: evolutionary game theory.

2.2 Cultural Evolution

Philosophers use EGT to model cultural evolution, which is change of belief over time. As such, it can be a tool for inquiries into the emergence of moral norms. Alexander is not the first to use EGT in this manner and such approaches share certain characteristics: Firstly, morality is posited as being an unintended emergence from repeated interactions between small groups of agents. It is thus a solution for frequently encountered social problems. Secondly, EGT does not model agents as being perfectly rational. This makes EGT a fitting method for Alexander's purposes: It is a study of boundedly rational agents engaged in repeated games. EGT models represent cultural evolution in game theoretic terms. Each model grounds on three element: a representation of the population, specific dynamical laws governing changes in the state of the population and games to model the interactions between agents. Individuals repeatedly interact each generation, employing strategies to maximize preference satisfaction. They collect payoffs according to the game's payoff matrix. Each player will then periodically review and potentially update her strategy, according to a learning rule. Over time, the frequency of individually successful strategies will increase in the population, while the unsuccessful go extinct. Strategies dominating the population over time prove their evolutionary

[1] For a detailed explanation of the experimental violation of the Sure Thing Principle, see Alexander, 2007 p. 11.

advantageousness. Alexander explores an array of different EGT models. The most realistic is Skyrms's and Pemantle's dynamic network.

3 The Dynamic Network Model

The dynamic network model (Skyrms and Pemantle 2000) is especially realistic because it includes both strategic and structural dynamics. Like all EGT models, it grounds on a representation of the population.

3.1 Representation: An Agent-Based Model

This model uses an agent-based representation, keeping track of the agent's identities even when they switch strategies or spatial position. It contains a finite population of agents, as well as a set of constraints on the possible pairings between them. Such constraints function as the social network. In Alexander's models, the possible interactions between the agents are visualized as a set of edges *E* on an undirected graph. All individuals connected to an agent represent the agent's neighborhood. Adding structural and strategic dynamics will allow for changes in strategy frequencies, as well as for the evolution of neighborhoods.

3.2 Strategic Dynamics

The strategic dynamics govern how agents update their strategies: In every generation each agent may interact (i.e. play the game) with some players connected to her. The agent reaps a total score consisting of the combined payoffs from each game. At the end of a generation, a player may review her strategy: She chooses some subset of her neighbors to compare her score to. According to a learning rule, she may switch strategies. Alexander limits the agent to the simple "imitate the best neighbor" heuristic, according to which the agent adapts the strategy of the highest scoring observed neighbor.[2] This mechanism drives strategy frequencies to change within a population. Alexander introduces a global parameter specifying the pace at

[2] If two or more observed neighbors tie for first place, the agent breaks the tie by rolling a die, weighted according to the number of tied winners.

which the strategic evolution occurs: p_s, "the probability that a given agent will adjust her strategy at the end of any generation."(Alexander 2007, p. 52)

3.3 The Structural Dynamics

Alexander turns to Skyrms's and Pemantle's paper (2000) to define the structural dynamics by which the social network evolves. In the dynamic network model, social structure is not defined before the dynamics start working. The network emerges as a product of the dynamics. Initially, all pairwise interactions are equally likely - the model is unstructured. Each agent interacts with a player at random. The game's payoffs then reinforce some interactions and make others less likely. In the authors' words: "The network structure emerges as a consequence of the dynamics of the agents' learning behav- ior" (Skyrms and Pemantle 2000, p. 01). The model's agents adjust the social distances between one another, which are represented by interaction probabilities determined by weights. Each agent *i* has a vector of weights ($w_{i1}, \ldots w_{in}$) that she assigns to other players. These weights compute the respective interaction probabilities (2000, p. 02):

$$Prob(agent{-}i{-}visits{-}j) = w_{ij} / \sum_k w_{ik} \quad (1)$$

In the beginning, all weights are 1. The initial probability of any two agents interacting is $1/(N - 1)$, with N as the population size. If an agent chooses to readjust her interaction probabilities to the encountered player, she simply adds her received payoff to the initial weight. A social distance matrix contains the interaction probabilities for all possible pairings. In Alexander's visualization, the network is represented as the set of edges, modified twofold: firstly, all possible edges are present, yet more or less prominent. [3]Secondly, the edges are directed: interactions between agents are more akin to visits, as only the party initiating the interaction receives a payoff.

[3] Those interactions, which are likelier than others are represented by more visible graphs: Edges representing an interaction probability of 1 being completely black, 0 meaning a white edge.

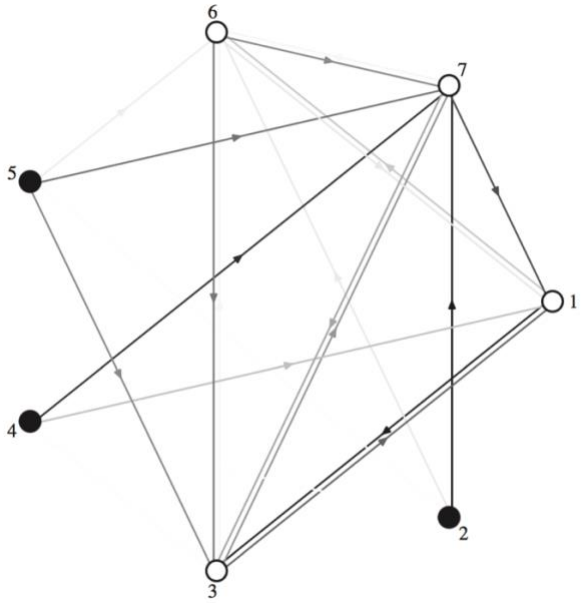

Figure 1: Dynamic Network after 1000 generations, without strategic dynamics. Payoffs: T = 4, R = 3, P = 2, S = 1 (Alexander 2007, p. 92)

Here too, Alexander introduces a parameter to specify the speed at which the structural evolution works: p_e, "the probability that a given agent will adjust her edge weights at the end of any generation" (Alexander 2007, p. 52). Both the social and structural evolutions crucially depend on the game the agents play and the payoffs they collect. For Alexander's purposes, this will be the standard Prisoner's Dilemma.

4 Universal Cooperation

Alexander uses the dynamic network to model the strategy of cooperation taking over a population trapped in a Hobbesian state of nature. He concludes that society's structure is sufficient for preventing universal defection in a population of self-interested agents, thus showing that our cooperative moral norm matches with what makes everyone best off.

4.1 The State of Nature

In *Leviathan*, Thomas Hobbes argues that without an all-powerful head of state, man in his initial condition could never cooperate. In this so-called state of nature, people are equally strong and resources are limited. All possessions are coveted and thus, when interacting with

others we may either attack them preemptively in order to gain their resources or lie low. With everybody preemptively attacking, society would devolve into a war of all against all, life being "evil, nasty, brutish and short" for those undergoing it. Game theorists model the interactions in such situations as Prisoner's Dilemma (PD), where agents may either cooperate or defect. These strategies correspond to Hobbes' preemptive attack and lying low. In its simplest form, the PD is described by a payoff matrix of this form (Kuhn 2019):

	Cooperate	Defect
Cooperate (Lie Low)	R,R	S,T
Defect (Anticipate)	T,S	P,P

The payoffs must follow this chain of inequalities:

$$T > R > P > S^4 \qquad (2)$$

Mutual cooperation is a pareto-optimal outcome. However (D,D) is the game's only strong Nash equilibrium. For both players, defecting is the strictly dominant strategy. Therefore, only "irrational" players would mutually cooperate and earn the highest reward. This holds for standard rationality assumptions in one-shot games. However, the situation changes when the game is re-iterated and social structure emerges. In the following section, I explain how Alexander uses the dynamic network to model dominating cooperation from an initial state of nature.

4.2 Cooperation Spreads

Alexander shows how groups of cooperators eventually dominate the model's population. However, in order to model this course of evolution, some assumptions about the initial state of the model must be made.

1. Agents must be free to visit more than one opponent each generation.

[4] R is the reward both players receive when they mutually cooperate. P signifies the punishment when both defect. T ("temptation") represents the reward a player reaps when she defects when her opponent trusts her. The betrayed player gets S, the sucker's payoff.

2. *pe* > *ps*: The structural dynamics must move considerably faster than the strategic dynamics. Agents must be locked in interactive clusters before starting to review their strategies.
3. The clusters must be large enough, relative to the payoff matrix of the PD. They must exceed T/R members (Alexander 2007, p. 99).

Lastly, note that only the agent initiating the interaction receives a payoff (2007, p. 94). The visited player's generational payoff is not affected. With these starting assumptions in mind, I now sketch the cooperative development of Alexander's model (as discussed in 2007, pp. 94-100).

If the strategic dynamics are held constant (i.e. $p_s = 0$) the population will aggregate into two types of clusters: cooperative clusters (C,C) and exploitative clusters (D,C) will crystalize. This preferential association is obvious: When a cooperator visits a fellow cooperator, she receives payoff R, and if she plays against a defector she is punished with S, the sucker's payoff. When she readjusts her weights, she will decrease her social distance to cooperators and reduce her interaction probability with defectors. Similarly, the payoff structure of the PD-matrix discourages the defector from visiting fellow defectors (as it means earning P), while strongly encouraging him to visit cooperators, as it means earning T (the highest payoff). This cluster formation does not yet consider the possibility of strategy updates.

Alexander shows that this structural development can be expected, no matter how the PD-matrix is constructed. In his simulations, after 1000 generations, the chances of a cooperator interacting with a defector are nearly zero (2007, p. 96). Given the clusters are in place, if the modeler now activates the strategic dynamics $p_s > 0$, then cooperation will "always dominate in the limit" (2007, p. 52). However, one more "modest condition" is necessary. The latter refers to 3: The cooperative clusters must be larger than T/R. Here is why: Strategy switches can only occur via the "imitate-the- best" heuristic. After being locked in a cooperative cluster, cooperators will (virtually) never choose to interact with defectors. As they only encounter fellow cooperators, at every strategic review, they can only imitate their kin. Thus, cooperators never change. They are locked-in – both structurally and strategically.

This does not hold for defectors, who exclusively interact with cooperators within exploitative clusters. For a defector to switch, it only takes one generation in which the defector reviews his strategy when the highest-scoring neighbor is a cooperator. This can only happen if 1. multiple

visits are allowed and 3. the cooperative cluster is large enough: During the switching round, the defector must only interact with one cooperator within his cluster, thus reaping T. For the switch, it takes a cooperator within the cluster to earn a score higher than T. This is possible if she interacts with every cooperator within her cluster C. She then earns a score of $C \cdot R$. If $C \cdot R > T$, i.e. $C > T/R$ holds, the defector will imitate the highest-scoring cooperator within his cluster. He will have become a cooperator.

Given 1 – 3, there is a non-zero switching probability for all defectors. Therefore, in the long run, cooperation will dominate. However, the likelihood of early cluster formation (and thus of universal cooperation) negatively correlates with p_s. Using a simulation Alexander shows: "As p_s approaches zero, the probability of the world arriving at a state of universal cooperation sharply increases. For $p_s = 0.0125$, universal cooperation occurred over 95 percent of the time" (2007, p. 100). Note that throughout this mechanism the structural dynamics have been fixed at $p_e = 1$ (2007, p. 52).

4.3 Alexander's Conclusion

The relation between morality and rationality seems to hold: Our cooperative norm has proven evolutionarily advantageous in the model. As Alexander set out to show, society's structure (the clusters) brought about universal cooperation. Alexander uses the model to infer that "the structure of society may prevent the war of all against all as effectively as the sword" (2007, p. 100). In the following section I evaluate the model and challenge Alexander's inference.

5 Evaluation

Without commentary, any model is merely an imaginary world. In this evaluation, I discuss how Alexander uses the dynamic network to explain our real moral norm and assess whether it grounds inferences to true conclusions about the real world.

Firstly, I discuss what exactly Alexander's model is aiming for: Does it aim for a historical account of the emergence of our morality? Or is the model's purpose less grand? Alexander sends mixed messages about the model's explanatory aim. I argue that the model aims to show that cooperative norms are evolutionarily advantageous, thus exposing the relation between rationality and morality.

Secondly, I assess whether the model has succeeded in exposing this relation. The model aims to show the evolutionary advantageousness of our real moral norm. To that end, the model must grant epistemic access to the real world. I use d'Arms' et al.'s guidelines to assess whether the model permits model-to-world inferences.

5.1 The Explanatory Power of the Model

What exactly is Alexander arguing for with the model? In the book he sends mixed messages about his intent. In the preface Alexander foreshadows:

"This book argues for the claim that much of the behavior we view as "moral" exists because acting in that way benefits each of us to the greatest extent possible (. . .)" (2007, Preface)

This is an ambitious claim, which I call the strong inference:

- The model[5] shows the reason why, and the actual mechanism by which our morality evolved.

This is not what Alexander argues for throughout his book. Rather, he aims to make a weak inference:

- The model shows that our moral norm would prove similarly advantageous in real evolving populations under similar conditions. Therefore, "the appropriate relation between acting in accordance with that norm and long-term maximization of expected "utility" holds"(Alexander 2007, p.33).

The weak inference is necessary for the strong inference: If the model aims to explain the actual emergence of morality, then it must show that the norms would prove evolutionarily advantageous in the real world. Whichever way Alexander puts it in text, the model cannot possibly ground the strong inference. As Uskali Ma ̈ki points out, many models do not set out to answer "why" questions. They are not meant to explain phenomena by indicating how they really happened. Alexander's model is one such model. Rather, the model addresses this type

[5] For reasons of simplicity, I will use "the model" to refer to our moral norms proving to be evolutionarily advantageous in the model.

of question: "How could P have come about?" (Maki 2009, p. 24) If applied successfully, the model permits an inference to one possible explanation, by exposing one potential mechanism that could have produced the observed outcome.

Therefore, the reader looking for an empirically accurate answer to the question "Why do our moral theories look this way?" will find no answer in Alexander's book. However, the model answers the question: How could our moral theories have come about? Alexander proposes a modest answer to this in his book: "Boundedly rational agents tend to promote moral behavior. (. . .) it does seem that it holds often enough for it to be more than a mere coincidence." (Alexander 2007, p. 267)

Why can the model not answer "why" questions? It can isolate only a subset of many working mechanisms driving real human evolution, rather than exhaust all possible explanations. Such criticism is leveled at EGT in general: Its explanations do not produce precise descriptions of the mechanisms driving the evolution producing the examined behavior. One is left to wonder how much explanation can be found in the strategic aspect of evolution compared to other mechanisms. The evolution of any phenomenon is a unique historical event and can only fully be discovered empirically. EGT may exclude certain possible historical sequences. However, it cannot indicate that a unique historical sequence of events suffices to bring about the explanandum (Alexander 2019).

Therefore, it would be wrong to assume that the aim of the model is to explain the etiology of our morality. Alexander's model can however give one possible answer as to why we hold our moral theories: They serve as evolutionarily advantageous heuristics. This conclusion would by no means be inconsequential for our understanding of morality: It could serve as an explanation for the persistence of our moral norms, as well as provide a reason for their normativity.

However, in order to make this point, Alexander has to do more than show that cooperation can emerge in the model. He also needs to infer that cooperation would prove similarly advantageous in real populations as well. Does his model allow for such inferences?

5.2 The Weak Inference

In this section, I assess whether Alexander's model provides grounds for inferences about the real world. Does the model show that our moral norms would prove similarly advantageous in real populations?

5.2.1 M¨aki's Framework

I introduce Uskali Ma¨ki's framework in order to structure my evaluation. In his paper *MISSing the World* (2009), M¨aki provides a useful way of thinking about models in general: Good models function as surrogate systems.

Models are used to gain epistemic access to complex target systems. To represent a target is to be its surrogate system. If successful, the surrogate allows for inferences to true conclusions about the target (M¨aki 2009, p. 41). Permitting such inferences requires a minimal degree of resemblance: The mechanism in the surrogate system must be analogous to the observed mechanism in the target system. In M¨aki's terms, the model must settle issues of resemblance to the target favorably. If so, then the model grounds inductive inferences to the target. It is a credible surrogate system, allowing for model-to-world inferences. It is a bridge to the real world (2009, p. 35).

How does Alexander's model fare in this framework? The dynamic network model tries to explain how real populations would evolve in situation akin to Hobbes' initial condition. To that end, the target system is a real evolving population under similar conditions. I refer to this target as the environment of evolutionary adaptation (EEA), using D'Arms' et al.'s terminology (D'Arms et al. 1998, p. 86). Such a population need not exist at the moment. Rather, it is important that the model raises issues of resemblance with a realistic conception of how real individuals would interact in such circumstance.

Because the model aims to represent the EEA, issues of resemblance arise. However, are they settled favorably?

5.2.2 Does the Model Bridge to the World?

Alexander can make the weak inference only if the model is a credible surrogate for the EEA. To this end, the modeled mechanism must be sufficiently analogous to the mechanism of real cultural evolution. I use D'Arms', Batterman's and G´orny's guidelines for the assessment of EGT explanations of human behavior to assess whether the model sufficiently resembles its target. The guideline demands a model be representative, robust, and flexible.

a. *Representativeness. Circumstances with the structure of the math- ematically characterized interaction which the model treats must be realized with sufficient frequency in the EEA.* (D'Arms et al. 1998, p. 89)

Here, I examine the structural representation the model uses.[6] Do the modeled interactions in the dynamic network sufficiently correspond to real interdependent problems in real populations?

The network's directed edges are problematic: Only the player initiating the "visit" receives a payoff. The structural dynamics make for exploitative clusters. Their exploitations are victimless: The defector is rewarded with the highest payoff T, the cooperator does not get punished with the sucker's payoff. This is an unorthodox way of using the PD in repeated interactions: The defector is constantly reaffirmed in his strategy, while the exploited cooperator remains unaffected. I can think of no real, frequently occurring scenario of this type and Alexander does not argue for this idealization. The concept of victimless exploitation does not cohere with the Hobbesian state of nature either. Rather, in this description of the initial condition, Hobbes claims all resources to be scarce and necessary for survival. If an individual falls victim to unsuspected defection (i.e. preemptive attack) she will most certainly feel her powers decreased measurably. Because this is not the case in the model, it does not sufficiently represent the EEA's interactions.

b. *Robustness. The desired result (here: universal cooperation in the limit) is achieved across a variety of different starting conditions and/or parameters.* (D'Arms et al. 1998, p. 90)

In his modeling Alexander works backwards: From the desired conclusion he reverse-engineers the favorable initial conditions for the model. To that end, he makes these assumptions:

1. Agents must be free to visit more than one opponent each generation.

[6] It would also be called for to assess the payoff matrix of the game: Only if circumstances with the structure of a PD have been a frequent part of human life can the success of cooperation in the model explain that it would prove successful for real populations. Undoubtedly, there is a consensus among philosophers that the PD is appropriate to model interaction in states of nature. I will not challenge it here. The bigger issue lies elsewhere.

2. The structural dynamics must move considerably faster than the strategic dynamics: $p_e > p_s$

Before assessing the assumptions, I note that the likelihood of the desired result remains unaffected by changes in the values of the PD-matrix. Alexander shows this (Alexander 2007, p.97). Nevertheless, the conclusion is structurally fragile. Firstly, assumption 1 is necessary: "when only one interaction is allowed, cooperation will not dominate in the limit"(Alexander 2007, p.99). Because both review processes can only take place at generation's end, agents visited more than once during a round cannot readjust their weights or strategies between the interactions. This is not a strong assumption but it is structurally necessary.

By virtue of its agent-based representation, the model is undetermined. The likelihood of universal cooperation negatively correlates with p_s. Alexander uses computer simulations to calculate this likelihood, given some regions for p_s. He only reveals the results for values 0.1, 0.05, 0.025 and 0.0125 and merely mentions the most favorable result in the text. However, for $p_s = 0.1$, the likelihood of eventual universal cooperation falls to 0.575.[7] This hints at the instability of the result: With a smaller contrast between p_s and p_e, the likelihood of early cluster formation shrinks, making universal cooperation improbable. Can this contrast reasonably be assumed? Low values for p_s model reluctance to strategic change. As such, this is not implausible: switching strategies might come at great cost. However, so does re-adjusting one's relationships. Remember that p_e is fixed at 1. An agent will adjust her social surrounding every generation, whereas in Alexander's proposed scenario ($p_s = 0.0125$) she will only review her strategy around every eighty generations. This implies that changing one's social surroundings is always considerably less costly than even trying a different strategy. Why think that individuals are especially unwilling to try new strategies, while being very flexible socially? Alexander needs to give a justification for this assumption. He provides none. The only argument for choosing this initial condition is its necessity to engineer the desired result. A less controversial, more nuanced assumption about p_s and p_e would render the chances of universal cooperation too small to make Alexander's point. It is safe to assume that $p_s = 0.5$ and $p_e = 1$ would not sufficiently favor the result. The desired result is structurally fragile. The initial

[7] For a visualization of these results, see figure 3.34 on Alexander, 2007 p. 100.

assumptions are demanding and only a small subset of all possible initial conditions can compute the desired result with sufficient likelihood.

c. *Flexibility. (i) The evolutionary strategy whose adaptiveness the model demonstrates is potentially realizable by a number of different mechanisms. (ii) The model itself can be understood to represent different possible processes.* (D'Arms et al. 1998, p.91)

(i) Here, I examine the mechanism by which cooperation spreads in the model. Governing this mechanism is the singular learning rule "imitate- the-best-neighbor". In Alexander's model, this imitative heuristic is crucial for cluster formation and brings about the desired result. Therefore, in this model, the universal cooperation rests on restricting the agents to this heuristic. Cooperation's adaptiveness can only be realized via one mechanism: (i) does not hold for the model. This is problematic. As D'Arms et al. point out: "The more seriously we want to take the idea that the players in these games are rational agents, the more we should look for ways of learning from the environment other than simply imitating the strategies of those nearby."(D'Arms et al. 1998, p.94)

If game theory's rationality is too demanding to model real people, then Alexander's representation is too simplistic to capture them as rational. The model dramatically caps the individuals' rationality: Whether or not it is rational to imitate or to choose a different strategy (to anti-correlate) depends on one's social surroundings. However, the model's agents are not free to anti-correlate. Alexander's restriction is a controversial assumption, especially when one considers the agents' situation: For realistic cooperators, would it not be tempting to "mutate" to defection, taking advantage of one's location within a cooperative cluster? When the game is reiterated, it becomes especially unlikely that realistic cooperators would remain locked-in, unfazed by this temptation. Mutation by temptation of this kind coheres with Alexander's bounded rationality: Although only satisficing and cognitively limited, the agents remain self-interested, acting solely to maximize their individual expected utility. Given the PD-Matrix there is ample opportunity cost in remaining a cooperator within a cooperative cluster. Therefore, Alexander's choice to restrict agents to the mimicking learning rule fails to capture agents as appropriately rational. However, cooperation is only realizable via restriction to that single learning rule. It appears that the modeled universal cooperation rests on a subpar representation of rational agency – bounded it may be.

(ii) Lastly, the model can nevertheless be understood as representing different possible processes. Specifically, it can be understood as representing any cultural evolution within conditions similar to Hobbes' state of nature (i.e. scarcity of resources, no civil authority etc.).

5.3 Critical Reflection

Alexander's model falls short of most of D'Arms' et al.'s criteria. I used these to answer whether the model settles issues of resemblance to the target system favorably. It does not: The mechanism of cultural evolution in the model is too unlike the mechanism driving realistic cultural evolution. This suggests that Alexander's attempt at showing that cooperation spreads is a weak failure (using Ma ̈ki's terminology): The model longs for epistemic access but fails to resemble its target in appropriate ways (Ma ̈ki 2009, p. 36). Therefore, it cannot bridge to the world via model-to-world-inferences: Alexander cannot credibly infer the evolutionary advantage of cooperation from the model to our real norm. This does not mean that the model is without merit, especially when appreciated in its context. In the book, Alexander employs an array of increasingly more realistic models. The realism grounds in the integration of structure.

Although realism is not always an improvement on the explanatory power of a model, it is here: As surrogate systems, EGT models aim to show whether strategies are advantageous in real cultures. To that end, they need to sufficiently resemble the target system – they must be sufficiently realistic. If the dynamic network model is unfit to represent real cultural evolution, then aggregative or less complex agent-based models do not stand a chance.

Alexander has shown that the more realistic a model is, i.e. the closer the model resembles our evolution, the more likely it is that purely self- interested agents will choose our norms as the most effective strategy. This suggests that with increasing realism, the models carry greater explanatory power. There is no reason why EGT cannot explain the facets of our morality. The evolution of these models is by no means complete: Integrating structural dynamics is only one of many ways of making EGT models more lifelike. Sets of possible strategies can be added, the global parameters (p_S and p_e) may be adjusted over time (representing the increase of switching cost over time). An exploration of possible improvements is, however, outside the scope of this paper.

6 Conclusion

In this essay, I have explored Alexander's *The Structural Evolution of Morality*. Specifically, I have reproduced and assessed his usage of Skyrms' and Pemantle's dynamic network model. Alexander shows that within this model, cooperation may emerge and ultimately dominate a population trapped within a Hobbesian state of nature. He claims this to show that our real cooperative norm is in fact evolutionarily advantageous. I have challenged this inference, drawing on Ma¨ki and D'Arms et al. The model does not sufficiently resemble real cultural evolution, so as to allow for the inference. Put in M¨aki's terms: the surrogate does not bridge to the world. This model nevertheless is a considerable improvement on its predecessors. Within the context of the book, the dynamic network model appears as part of an overarching development: EGT models are becoming increasingly more realistic, thus becoming better tools for inquiries into cultural evolution.

References

Alexander, J. M. (2007). *The Structural Evolution of Morality*. Cambridge University Press.

Alexander, J. M. (2019). Evolutionary Game Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019 ed.). Metaphysics Research Lab, Stanford University.

D'Arms, J., R. Batterman, and K. Gorny (1998). Game Theoretic Explanations and the Evolution of Justice. *Philosophy of Science* 65 (1), 76–102.

Kuhn, S. (2019). Prisoner's Dilemma. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University.

Ma¨ki, U. (2009). MISSing the World: Models as Isolations, Representations, and Credible Worlds. *Erkenntnis* 70(1), 29–43.

Skyrms, B. and R. Pemantle (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the United States of America* 97 16, 9340-6.

Verbeek, B. and C. Morris (2018). Game Theory and Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 ed.). Metaphysics Research Lab, Stanford University.