A Morality Evolutionary Game Theory Can Model

Mikhail Volkov

Department of Philosophy, Logic and Scientific Method
London School of Economics

Abstract

Evolutionary game-theoretic (EGT) models of morality face powerful under-addressed objections. Critics claim the simulations fail to specify their explanandum, making their explanatory value murky. Additionally, morality is suggested to be a concept not computationally representable, jeopardising the method's general applicability. This paper explicates and addresses the objections. I argue that at least one concrete conception of morality, epistemic emotionism, can be a plausible subject of EGT explanations. I analyse how fixing this explanandum assuages the methodological objections and provide a computational model as proof of concept. If successful, the contribution placates serious long-standing criticisms of EGT as a meta-ethical tool.

Wordcount: 8925 (excl. bibliography)

1 Introduction

We behave morally even if doing so goes against instrumental rationality in local cases. For any given visit to a new cafe, it profits me to eat and run away without paying, also to steal when I can be sure no punishment will ensue – and yet I do not. Why?

Attempts to explain moral behaviour in light of its occasionally instrumentally counter-productive nature are not new. They figure centrally in the moral philosophy of Hume, who famously supposed humans unable to satisfy their needs and passions but through 'mutual succour'. Conventions that demand parties do their share in a mutually beneficial enterprise simply are the rules of justice, according to Hume. This mutually beneficial nature of conventions is recognised and motivates practice: 'When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behaviour' (Hume 1896, Book III, Part II, Section III). In turn, human sense of sympathy with the public interest is responsible for our feeling moral approbation or disapprobation towards those who break or abide by said conventions (Book III, Part II, Section II): '[T]he sense of moral good and evil follows upon justice and injustice.'

More recently, philosophers – some explicitly espousing Hume's project – have adopted game-theoretic and evolutionary game-theoretic methods to vindicate the emergence of justice and morality as beneficial products of social learning and mimicry in the context of repeated, socially situated interactions. Different formal techniques have allowed modellers to demonstrate that in a range of scenarios modelled as non-cooperative games, strategies mirroring moral

behaviour can receive wide and stable uptake in communities if agents are guided by utility-oriented rules for strategy change. These results, so goes the argument, might help elucidate the origin of justice and the social contract (e.g. Skyrms 1996; Bruner 2018), of morality or moral sense (e.g. Alexander 2007; Bruner 2021) and other similarly thick concepts characterising social behaviour.

The most salient philosophical question is whether these models manage to adequately account for the origin of something that can plausibly be called morality or justice. Powerful criticisms have been repeatedly advanced, arguing that the models do not (Levy 2011; Kitcher 1999; D'Arms 2000) or, more radically, can never do so (Alexander 2007, Ch. 8, Arnold 2008). While modelling the emergence of morality has since continued, the critiques have not received a convincing response to the effect that EGT can capture a thick explanandum like morality adequately and in a philosophically useful manner. This work attempts to do exactly that. Critics argue that thick moral phenomena are often unspecified as an explanandum and, if given adequate definition, EGT cannot capture them. This paper provides a response to the critics and shows that at least one existing meta-ethical conception of morality can receive a translation into computational models. There exists at least one independently motivated understanding of morality for which EGT may assist the evolutionary explanation. The critics' blanket scepticism is thus shown wrong by means of a counter-example. As such, it is possible more moral theories apt for interesting EGT analyses exist.

The context of previous EGT work is important for the current contribution. Section 2 surveys the relevant EGT works and sketches the logical structure of explanation implied by them. Section 3 outlines the criticisms against the method's explanatory value (from then onwards, I focus specifically on explanations of *morality*, as have the critics). After Section 3, we will have been left with the following upshot: EGT explanations of morality leave their explanandum unspecified *and* we have reasons to be sceptical of the method's capacity to provide an adequate account of a phenomenon as rich as morality, regardless of how it is specified. Section 4 offers the critics a response and a positive proposal for modellers of morality. I argue that epistemic emotionism as formulated in the meta-ethical literature coupled with a functionalist conception of emotions is an apt disambiguation of morality for EGT to attempt explaining. Section 5 further sketches how this concept can be cashed out computationally and shows that its computational representation may serve as an adequate response to the sceptics. Section 6 concludes.

If successful, this work defends evolutionary game-theoretic modelling from explanatory irrelevance (a frequent concern for the method's practitioner), and connects formal modelling work with debates in value theory. Evolutionary game-theoretic work can be directly relevant for meta-ethical discussion – but then, a more careful treatment of its explananda is needed.

2 Evolutionary Game-theoretic Models of Normative Phenomena

EGT subsumes formal methods centred around boundedly rational populations of agents interacting via games. As applied to modelling dynamic processes, these include variant continuous population dynamics, discrete-time models like the Moran process and, more recently, agent-based local interaction models. In addition to this variance in methods, EGT has been applied to distinct normative phenomena such as morality, cooperation, and just social contracts based on fair bargaining. Due to the underspecification of the explananda, whether an EGT model reflects morality or aspects of the social contract is often up to the modeller's own interpretation rather

than something inherent model features. Convergence to cooperation in a Prisoner's Dilemma can support the emergence of trust, a social contract or cooperation as a general phenomenon. It is hence useful to consider EGT applications to various social phenomena as even those models that do not explicitly concern morality share explanatory structure with those that do and may be interpreted as such. Additionally, modellers working with thick normative phenomena of social behaviour such as justice may be facing the same objections as will follow shortly concerning morality in particular.

Skyrms' modelling of the evolution of justice starting with (1996) has been foundational in the philosophical employment of EGT. Using replicator dynamics, Skyrms models several games and structural assumptions that support the evolution of 'just' outcomes in them. Replicator dynamics makes the success of each strategy vis-a-vis the population average the determining factor in how widely the strategy spreads – it is a qualitatively adaptive dynamics (Skyrms 2000). Informally, this says that individuals who do worse than the average of the population are forced to change strategies to those that outperform average. The explanatory strategy goes like this. Replicator dynamics is applied to a game mirroring an interaction where some norm of justice may manifest, like the Nash demand game in figure 1 modelling resource division. Then, conditions enabling population-wide convergence to a strategy antecedently defined as moral are analysed. In figure 1, action profile (fair, fair) naturally constitutes the just (or moral, or fair) outcome: agents split the windfall evenly.

		Player 2		
		Demand 3 (meek)	Demand 5 (fair)	Demand 7 (greedy)
Player 1	Demand 3 (meek)	**3, 3**	**3, 5**	**3, 7**
	Demand 5 (fair)	**5, 3**	**5, 5**	**0, 0**
	Demand 7 (greedy)	**7, 3**	**0, 0**	**0, 0**

Figure 1: Divide-the-Cake (DtC)

In orthodox game theory, (fair, fair) is one of three equally viable pure Nash equilibria. It thus fails to explain the fair outcome's real world salience and our special attitudes towards it. One can hope that embedding the game into a dynamic, social setting of an EGT model will yield a better explanation.

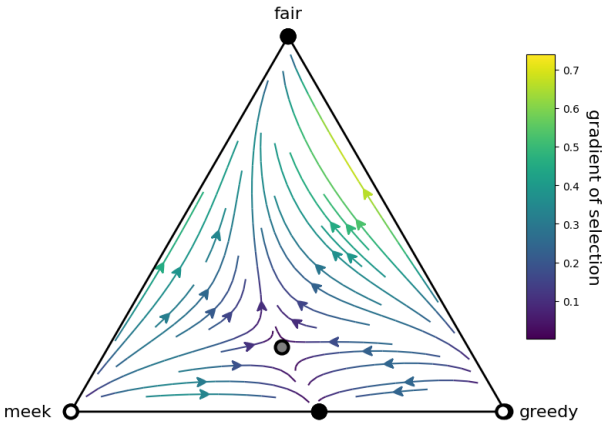


Figure 2: Simplex of Replicator Dynamics of Divide-the-Cake

Indeed, from most initial distributions, the population ends at uniform fair sharing (figure 2). We thus have a partial explanation for the salience of fair sharing: it is the strategy the population is likely to adopt if strategy change is qualitatively adaptive. Alongside fair division, Skyrms models further demands of justice within EGT framework, including punishments for unfair offers in the Ultimatum Game and mutual cooperation in Prisoner's Dilemma that becomes sustainable under correlated interactions between proponents of same strategies. Explorations like these, it is claimed, constitute some progress towards explaining the origin of justice (Skyrms 1996, Ch. 1).

An alternative strand of modelling has employed the local interaction approach that more closely approximates real community structures (Alexander and Skyrms 1999; Alexander 2007; Skyrms 2003; Skyrms and Pemantle 2000). Given a network where agents (nodes) interact with their connections, each agent dynamically changes strategies according to a success-oriented revision rule. The latter is most commonly some form of imitating the most successful neighbour. Analysis proceeds by simulating the change of strategy frequencies in a population over many runs. E.g., tracking a network playing DtC, we again gauge that the uniform outcome is likely to evolve (figure 3). Modifications on such models include weighted networks of (Skyrms 2003, Part III) and (Skyrms and Pemantle 2000), where high-payoff interactions increasing edge weights, raising the probability of future interactions between agents.

Most comprehensively within this strand, Alexander (2007) provides local interaction simulations of different structures for a variety of games such as Prisoner's Dilemma, Stag Hunt, versions of DtC and others. Interestingly for us, Alexander couches his work in explaining *morality*. So, cooperative strategies in DtC become signs of the sense of *fairness*, in the Ultimatum Game of *retribution*, in Stag Hunt of *trust*, etc.

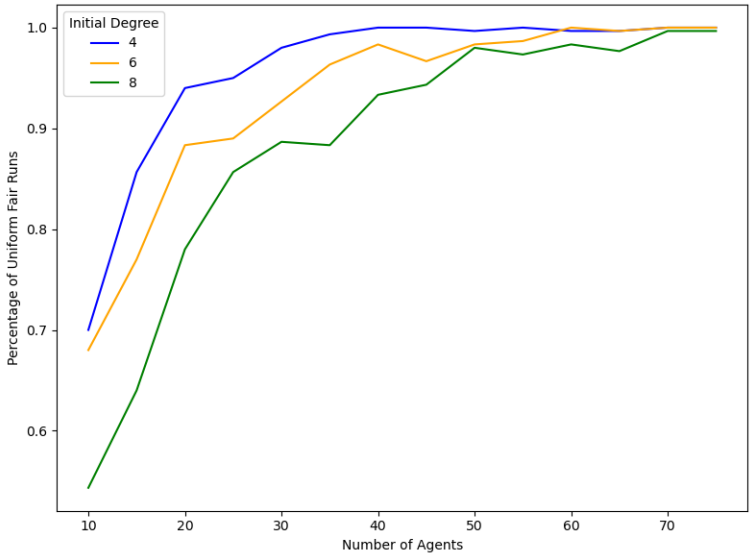


Figure 3: % of Uniform Fair Runs / Population Size for Watts-Strogatz Network Playing DtC

Recent EGT applications shy away from holistic explananda like morality and justice.[1] Nevertheless, they retain focus on similarly complex social phenomena like just bargaining schemes or cooperation.[2] For example, Bruner (2021) considers the Matthew-Luke game from (Braithwaite 1955) – a negotiation over playing time between two musicians who prefer to play given the other does not. Figure 4 displays a particular cardinal instance with symmetrical coordinated outcomes (generally, Braithwaite's problem does not require this assumption (Raiffa and Luce 1957, 6.11)). Which stable agreement will 'Matthews' reach with 'Lukes' after repeated interactions?

		Luke: Play	Luke: Don't Play
Matthew	Play	**0, 1**	**4, 3**
	Don't Play	**3, 4**	**2, 1**

Figure 4: Matthew-Luke (ML)

Applying two-population replicator dynamics to ML as a bargaining problem, Nash bargaining solution appears most likely to emerge.[3] For the game above, this suggests that populations gravitate to (Don't Play, Play) with 'Matthews' giving up their play time entirely.[4] Elsewhere (Bruner 2018), Bruner introduces metabargaining where negotiation over the feasible set precedes the play of bargaining strategies. Here, the utilitarian bargaining solution is privileged by the static EGT analysis using evolutionary stability. Since both Nash and utilitarian solutions are

1. Reasons for this shift are well put in (Bruner 2018): 'morality' as an explanandum has been tricky to pin down. This is the main issue this paper addresses.

2. For overview of the topic see (Maschler, Zamir, and Solan 2020, Ch. 16).

3. The details of Bruner's treatment are somewhat more involved, as he considers several bargaining setups many instantiations of ML's ordinal structure: Nash solution is favoured in some but not others.

4. Assuming (Play, Play) as the disagreement point.

recognised to be fair bargaining outcomes, these models are taken to demonstrate that fairness may emerge from boundedly rational populations in the process of repeated learning across a range of scenarios.

Generally, EGT work on bargaining has covered multiple intricate bargaining set-ups, games and assumptions (Vanderschraaf 2018; Zollman 2008). For instance, a replicator dynamics model from (Zollman 2008) shows how the evolution of fair division is sometimes likelier to arise in composite games that are combinations of two standard games than in individual games.

Our limited overview demonstrates that EGT modellers have tended to employ different methods and focus on slightly different normative explananda. Underlying all these models, however, is the following explanatory structure:

- In the real world, normative phenomenon X such as morality or justice is observed.
- In the real world, interactions transpire in a repeating social setting.
- In the model, *if* interactions transpire in a repeating social setting *and* agents are equipped with a success-oriented rule of strategy change based on the neighbour's and/or the average population success, they converge to behaviour aligning with the demands of justice and morality.
- Therefore, in the real world, normative phenomenon X has emerged due to success-oriented learning and imitation in the process of repeated interactions.[5]

EGT models based on success-oriented strategy change (which most are) will fall under a template resembling the one above. The upshot of this explanatory structure would be that we are fair, just and moral because we have learned, in a success-guided manner, from the rest of the population. Unfortunately, computational modellers generally do not flesh out the exact relation their computational demonstrations have to the philosophical phenomenon under discussion (Aydinonat, Reijula, and Ylikoski 2021). In our case too, there are multiple ways to interpret the relation between the two. I take the EGT modeller's ambition to be explaining the emergence of morality and justice (whatever the modeller implicitly meant by these terms) in a piecemeal fashion by modelling the emergence of different moral norms in varying scenarios. I also take the aim of the project to explain how the phenomena of morality and justice as these notions are employed in philosophical discourse have evolved in human communities. This is the *prima facie* reading of what EGT project is up to when setting 'morality' as its target. Namely, it attempts to uncover the causes of a holistic phenomenon, not pseudo-moral behaviour or an alternative history for how morality could have emerged, but real moral behaviour, with all its complexity unique to human sociality. This is what the language of authors like Skyrms and Alexander naturally suggests (Harms and Skyrms 2009, p. 1, Alexander 2007, p. 278).[6] The success of the project so conceived is at stake in the following discussion.

5. Space prevents me from showing exhaustively how widespread this implicit explanation actually is. The reader may also want to consult EGT work on the evolution of cooperation (Forber and Smead 2014; Harms 2000; Alexander 2015) and guilt and costly apology (O'Connor 2016; Rosenstock and O'Connor 2018).

6. This reading is sometimes met with the response that EGT does not owe a holistic explanation of how morality has evolved in any realistic detail but only the behavioural dimension of these phenomena.

I recognise the pull of this reading but contend that there are reasons not to adopt it. Firstly, as I show below, it undersells the potential of EGT models of morality. Secondly, the 'deflationary' version of the project needs further justification. For instance, we *know* that even the purely behavioural dimension of fairness is propped up by much more than prudential success-oriented learning – most importantly, also by internal motivations. It is not immediately obvious how a model-based explanation suggesting the opposite is of philosophical import.

3 Critique of EGT-based Explanations

Evolutionary game-theoretic explanations of morality in particular have come under extensive critique which can be divided into two families. One focuses on the empirical evaluation of models against our knowledge about the evolutionary environments where morality began emerging. Modelling assumptions are often found unsatisfactory in this respect (Kitcher 1999; Levy 2011). The other family concerns the nature of the explanandum and the relation that the models bear to it. This work gives a defence against this second critique, leaving the empirical evaluation and enhancement of the models for future work.[7] Within the second strand, there are two points of attack: (1) it is unclear what is meant by the word 'morality' in the modellers' works, and (2) there is no plausible sense of 'morality' that lends itself to computational translation. I consider the most powerful deliveries of the two concerns – from D'Arms' commentary on Skyrms' modelling of justice (D'Arms 2000) and from Alexander's critical reflection on his and his colleagues' explanations of moral norms (Alexander 2007), respectively.

3.1 Objection I: Explanations of *What*?

In articles that aim to provide EGT explanations of morality, fairness and just bargaining, one often finds that the explananda at stake are not given any explication or definition. Instead, it is silently implied that when a population in replicator dynamics or on a network widely adopts a strategy antecedently denoted as moral, explanatory success has been achieved. But precisely what has been so explained? Modellers cannot answer, insofar as they do not tend to specify what 'justice' and 'morality' mean in their work (D'Arms 2000). It is consequently impossible to assess these models as good or bad explanations of the phenomena in question: there is no clear target of explanation, so naturally there is no telling whether it is successfully reached.

This point is stressed by D'Arms (2000) and Kitcher (1999; 2014). While their responses concern Skyrms' work on justice, they apply to subsequent models of morality and just bargaining also. D'Arms starts with noting that the EGT project can be interpreted in two ways: as explaining the behavioural phenomenon of altruism or as explaining the distinctively human moral sense. The two are different in important respects. Altruism refers to a particular kind of behaviour, whereby agents act in favour of others and at a cost to oneself, and does not posit anything as such about the causes of these behavioural patterns. Clearly, morality is not a descriptive phenomenon in this manner: we normally think motivational patterns underlying our

7. While addressing the empirical critique in detail may fall outside the domain of a philosophy journal, let me briefly provide reasons why it is not obviously lethal to the project. With dynamic EGT models, empirical validation will bear mainly on structural assumptions such as ranges of parameters for the number of agents, topology, game specification, social hierarchies and on qualitative assumptions like the learning rules. For empirical validation, one must first decide on the species and period we take as the loci of moral evolution; here, *h. heidelbergensis* of middle Pleistocene are a point of some agreement (Tomasello 2016; Birch 2021; Sterelny 2021). One then accesses significant empirical fodder for models of human prehistory, including quantitative data on topology (Kelly 2013, pp. 171, 274; Gamble 2013, p. 73; Dunbar 1993, 2002), nature of group membership (Klein 2009, p. 81; Sterelny 2021, Ch. 2; MacDonald et al. 2021), group hierarchy (Fry 2006, p. 79; Sterelny 2021, Ch. 2), and inter-group warfare and interactions (Kissel and Kim 2018; Ferguson 2013a; Rodríguez et al. 2022; Weiss 1984, and Ferguson 2013b for overview on the topic). Targeted search for evidence becomes available for qualitative assumptions like the relevant evolutionary pressures of the time period (Gamble 2013, Ch. 2, 5; Tomasello 2016, pp. 44-45; Sterelny 2021, Ch. 1) and cognitive abilities of the hominin agents of the time (Harvati 2007; Thieme 1997).

In short, evidence on the origins of moral evolution that can receive direct computational translation is significant. If not outright completed, empirical validation of EGT models of morality can be improved significantly by turning to empirical literature.

actions to be a crucial component of our being moral. The presence of morality in a population must therefore necessitate additional facts about its agents apart from behaviour. Kitcher formulates this concern by saying there is a 'superstructure' of normative concepts to be accounted for (1999, p. 223). Worries about models missing crucial facts about human morality would survive even at the time of *The Ethical Project*, where Kitcher reiterates the critique (§8, 9). As examples of said additional facts, critics cite the existence of a system of punishments and internal feelings of guilt inherent in the members of the population (D'Arms 2000) or special evaluative attitudes towards other agents who behave in immoral ways (Kitcher 1999).

There is then a fork. We can take the models as showing the evolution of altruism and using 'morality' and 'justice' as unfortunate shorthands for it. But then, the models shed little light on the notions of morality and justice, unless the latter are understood in a non-substantive and philosophically uninteresting way. If we take the models to demonstrate the evolution of morality as a complex of distinctively human mechanisms for sustaining concrete behavioural patterns, they are unsuccessful because no such mechanisms are present or explained in the models. There is then simply no complex of things in the models that one can recognise as referents of the word 'morality'. We are forced to admit they fail to explain the presence of morality.

D'Arms notes a further complication: models disregarding additional mechanisms supporting morality in the real world are a *modus tollens* against any explanation of morality based on said models. This is because we know for a fact that real moral norms subsist on much besides prudential learning and behavioural changes. In contrast, these models entail that there is no 'need for a propensity for feelings of guilt when we 'unfairly' demand more than half – recognition of the lost returns should suffice to bring us back on track' and thus no difference between morality and expedience (D'Arms 2000). EGT not only leaves the thickness of morality and justice unexplained but speaks to the redundancy of the motivational patterns that accompany and propel real moral sense and norms.

The upshot: if modellers are agnostic about the meaning of their own explanandum, it is unclear whether their models are good or bad and what they are even models of. In turn, insofar the intended interpretation concerns the distinctively human tendency for moral behaviour, these models fail to explain it because they misrepresent central aspects of real moral behaviour.

3.2 Objection II: Pre-emptive Scepticism

Alexander's extension of the discussed objection (2007, Ch. 8) goes yet further and deserves closer scrutiny. D'Arms only targets the extant models of Skyrms and is overall optimistic about the promise of EGT models. In contrast, Alexander thinks the critique covers evolutionary game-theoretic modelling as such and thus *all* EGT models of morality. The scepticism is rooted in an alleged inherent mismatch between the method and the explanandum. All and only suitable notions of morality, goes the argument, are non-behavioural and contain motivational or superstructual components (mirroring our discussion in 3.1). For instance, we do not just punish and go about our day; we *want* to punish, enjoy when the wrongdoer receives just deserts and feel strongly that the punishment must happen. This combination of non-behavioural responses bears decisively on our action. Furthermore, *this* family of reactions seems to be what is truly interesting and puzzling about the evolution of morality as well as crucial to many meta-ethical views of it.

Alexander insists that EGT frameworks cannot adequately incorporate these superstructural components, as the dynamic of EGT is fundamentally behaviour-centred: payoffs accrue to strategies in the stage game and are the sole factor in the choice of strategy by the agent. In

(Alexander 2007, Ch. 8), the possibility of enriching the psychological make-up of modelled agents to more closely mirror the motivational structures at play in moral behaviour is entertained. Nevertheless, Alexander goes to conclude that the models are bound to underdetermine the crucial superstructural features of the agents' 'moral' actions:

> It doesn't matter that the strategy labels are 'punish' and 'enforce a norm', for the model still admits a purely behavioural interpretation...We don't want an account of evolutionary pressures that shows how people will come to act *as if* they are punishing defectors; we want an account of why people *really punish*. (p. 273)

This appeal needs precisification. As I read the point about the models invariably admitting a behavioural interpretation, it may be illustrated by an example familiar from philosophy of mind (Kirk 1974). Imagine a zombie world where humans behave just as we do in morally relevant cases but have no motivations underlying their behaviour. Perhaps as first thing in the morning they recite their strategy for the day:

Rule 1. 'Punishment strategy for unfair sharing'. If someone does not share evenly, punish them at the cost to oneself.

Rule 2. 'Boundedly rational dynamic for strategy change when sharing'. If I observe my social circle has stopped sharing but is doing better than I am, I stop sharing too.

and so on. Finally, they remind themselves that this is the recipe for maximising their success under constraints of social structure and their bounded rationality, after which they start their day. People in this world are decidedly not moral but mere expedient rule-followers. There is no morality in a world like this, *despite the inhabitants' behaviour matching that of moral agents*. However, the EGT models seem to be equally applicable to this world and the emergence of observed behaviour in the zombie world is as well explained by such models as is behaviour in our world. Exclusively focusing on its behavioural dimension, the models are thought to inevitably miss something crucial about morality, not to show 'why people *really* punish'. Since this behavioural reading can be applied to any EGT model of moral evolution, Alexander claims that for any thick notion of morality whose explanation is attempted, no adequate EGT representation can be had.

4 A Morality EGT Can Model

The criticisms certainly paint dark prospects: not only do *current* EGT models of morality not possess a clear explanandum and hence not explain morality, we should not expect *any* EGT models to do so, since they can only explain a behaviourist conception of morality which is not a plausible one. Surprisingly, despite the threat posed by these criticisms to the tradition of explaining normative phenomena in game-theoretic terms, few systematic responses have been forthcoming. This maybe at least in part due to some remaining ambiguity in the criticisms; it is unclear how they can be contested in a targeted manner. For instance, what would it mean to fix a conception of morality? What is the menu of such conceptions that the modeller can choose from? Further, it is claimed that agents in the models do punish, but not *really* punish like humans do. Maybe – but how do real humans really punish? It is thus implied that some consequence of there being real morality is missing from the models but we do not get anywhere near a precise indication of what that missing thing is, beyond a very broad gesture at 'a variety

of reasons and motivational structures' (Alexander 2007, p. 273) and 'superstructures' (Kitcher 1999). Thus, what exactly the critics want from the modeller still needs explication.

In this section, I propose a strategy that responds to both concerns outlined in the previous section on most plausible interpretations of what they are asking for. I further execute the strategy by providing a concrete and independently motivated notion of morality that accounts for our intuitions about its thick character (thus accommodating objections in 3.1) and give reasons to think it can receive a plausible computational translation (thus showing there is a counterexample to be had against objection in 3.2).

4.1 Explicating the Objections

Let us first consider more closely what philosophical work has to be done to offer a targeted response to objections. When D'Arms mentions the lack of specificity about morality as an explanandum, what may the modeller offer? Clearly, it has to be some notion of morality that does not deflate the meaning of the term into a group of agents simply *behaving* a certain way. It further has to provide a cluster of concepts with which the term 'morality' is bound for the purpose of the model-based explanation – that is, some meta-ethical commitment to what morality is.

But if not to its behavioural dimensions, then where else to look for the referent of the word 'morality'? Or: given a population of real agents, what would it take for us to conclude that these agents are not merely expedient but moral in the thick sense that we attribute to humans? From the preceding discussion in EGT literature, one inherits hints towards two other possible dimensions: the phenomenological and the functionalist (recall the critics' talk of 'motivational structures' (Alexander 2007, p. 273) and 'superstructures' (Kitcher 1999)). Let us see if a concept of morality that is computationally representable can be rooted in either of these dimensions.

First, the situations of moral significance in the real world are also accompanied by strong internal feelings. This phenomenological part is such a consistent presence in situations we label moral that it can be considered a necessary component for the notion of morality. The feelings of satisfaction when an amoral person gets punished, the feeling of guilt when we violate a norm. If the feelings are absent, then perhaps it can be argued that an important referent of 'morality' does not obtain.

Secondly, there is a functional or a dispositional role to morality that is distinct from either its behavioural manifestation or its phenomenology. That is, if a person is moral, then this must mean that necessarily, in situations of a certain structure, they feel compelled or motivated to act in a certain way. Importantly, this does not necessarily translate to behaviour: presumably, most humans have what it takes to be moral but do not always behave morally. This functional role of morality is just *a* causal influence on our behaviour, just a part of the equation. When we are in a Divide-the-Cake situation, it is wholly possible that some of us will not go for the fair split – but they will very probably feel some pull towards fair split or will suffer at least somewhat when they leave the game setting and look back on their choice. Furthermore, the presence of feelings cannot substitute this functional role either. The isolated phenomenology of feeling guilt – 'what-it-is-like-to-feel-guilty' – does not on its own have any relation to action. We may think it does because it is frequently, indeed almost inevitably, accompanied by the functional role of morality kicking in. When I feel guilty, I am at the same time much more likely to ask for an apology, I really want to do so. But it does make sense and will be useful for later discussion to divorce the feeling itself from the disposition to perform an act that we find to accompany the feeling in the real life. The separate functional role of morality is precisely this

being compelled or drawn to act morally and ultimately being more likely to act morally for it.

The critics' gripe highlights that the modellers make no attempt to capture any of these two extant components of morality that are arguably more important than the behavioural aspect when diagnosing a population with morality. Indeed, purely behavioural data may mislead about the moral status of persons and thus not necessary for a population being moral. Consider citizens forced by a dictatorial regime into participating in inhumane activities and snitching on their neighbours, despite feeling immense guilt and desire to stop while doing so. There seems to be a clear sense in which there *is* morality in this population, despite no behavioural manifestation being present.

Consequently, the modeller is well-advised by the critics to focus on capturing the other dimensions of morality when attempting to explain the latter. However, from the two other possible associations of the term 'morality', computational models cannot in principle track the emergence of phenomenological experiences. Indeed, it is unclear what this would mean: you cannot make computational units feel things and, in any case, you cannot make an observer recognise that they do. We are left with the possibility of tracking the emergence of the *motivational* dimension of morality. However, the way I read D'Arms, he is asking for a concrete meta-ethical commitment – a clear disambiguation of what morality is. The next subsection will be concerned with outlining such a conception rooted in the motivational dimension of morality. If the search for such a theory and its translation into a model possible, then there may exist EGT models to which one can in principle point and pronounce, 'by morality I mean what meta-ethical theory *M* means by morality, and the agents here are most plausibly understood to have evolved morality and not as-if-morality, where morality is understood as in *M*.'

4.2 Epistemic Emotionist Morality as the Proper EGT Target

To address the concern raised in 3.1, we need a more concretely formulated notion of morality such that it is clear what kind of morality our EGT models explain. In turn, if we want a meta-ethical theory that ties morality to some naturalistic motivational structures, a natural family of theories to consider are *emotionist* theories. Indeed, my claim, which I shall justify shortly, is that they form very suitable explananda for EGT models.

The choice of emotions-centred theory is salient when we appreciate that in psychology of emotions, the *functionalist* understanding of moral emotions is commonplace especially as it concerns their evolutionary function (Haidt 2003; Keltner and Gross 1999; Keltner and Haidt 1999; Hutcherson and Gross 2011) and that in recent EGT literature, the evolution of moral psychology has recently come into focus (O'Connor 2019, 2016).

The label of emotionism subsumes many metaethical views, of which Jesse Prinz' and Alan Gibbard's work are perhaps the most relevant examples (Prinz 2007, 2015; Gibbard 1992). However, many other authors, even of differing metaethical colours, admit the crucial role of emotions for a full-bodied concept of morality (e.g., Joyce 2005). Both Prinz and Gibbard align with the broad metaethical position that Prinz dubs epistemic emotionism (Prinz 2007, p. 16):

Epistemic Emotionism: Moral concepts are essentially tied to emotions.

Why think that moral emotions would in principle be easily translatable into EGT models? Emotions after all are things we feel – and the locus of morality as we identified it lies in motivation. This is where the functionalist view of emotions comes in. Let us take the two emotions that Gibbard considers crucial to his account: anger and guilt (Gibbard 1992; Clavien 2009). These picks are quite apt seeing how moral emotions are generally split into the categories

of reflexive (self-directed) and reactive (other-directed) emotions (e.g., Prinz 2007; Ben-Ze'ev 2000; Ellemers et al. 2019), of which guilt and anger respectively are perhaps most indicative.

The prevalent trend in the psychological literature on emotions is to define them as a function from an input to a behavioural output. For instance, in his influential classification of moral emotions, Haidt (2003) specifies each emotion by the elicitor and action tendency. In application to our emotions of interest, Ramsey and Deem (2022) cast guilt as a function that takes one's wrongdoing as an input and incentivises a self-detrimental signal aimed at triggering the sense of empathy in others. Tangney et al. (2013) and Vaish (2018) provide their own accounts of guilt, keeping closely to the functional understanding of the emotion. The same situation is observed when one consults the psychological literature on anger or indignation. Its output is taken to be the incentivisation of active interaction with others in order to correct their course of conduct (Hutcherson and Gross 2011; Haidt 2003, citations therein), whereby the correction involves some discomfort to both the punisher and the punished. Its input is similarly behavioural, even when it is understood differently by different authors: Hutcherson and Gross (2011) mention threat to oneself as the main trigger for anger, whereas Haidt (2003) takes goal blockage and recognition of unfair treatment to be the elicitor of anger. One can thus note that there is nothing too mystical about the workings of emotions as they concern behaviour in the psychological literature. They all operate with either structural ('goal blockage') or behavioural terms ('threatening action'). This functionalist reading of emotions would seem to lend itself readily to translation into computational vocabulary.

Hence, epistemic emotionism can serve as target for EGT modelling work if it is seeking a concept of morality to explain properly, not as-if. This is for several reasons. First: epistemic emotionism disambiguates the functional mechanism we expect to be active in a moral population in a computationally representable way. If emotions are understood functionally as they are in the science under whose domain the analysis of emotions falls, then they are functions from behaviour to behaviour. This is a representation of emotions that is encodable into a computational model. Second: it supplies a more well-defined concept of morality that lends itself to being further broken down into concrete components, i.e. concrete emotions. Thus, the functional role of each emotion can be tackled separately, with a corresponding empirical investigation into its evolutionary history and a suitable family of models representing its emergence. If this piecemeal modelling is successful, then these explanations can be joined into what on the epistemic emotionist picture would amount to an explanation of the fundamental moral concepts.[8] This sounds a lot like EGT delivering on its promise to explain morality.

5 Back to the Pre-emptive Challenge

We now have our eyes on the prize: emotionist morality with emotions defined functionally. Can this meta-ethical notion be cashed out computationally? The crux of the objection in 3.2 was precisely that similarly thick conceptions of moral behaviour cannot receive a computational translation because the latter forces association of morality with some strategy frequency.

While the present paper is mainly conceptual and programmatic, a proof of concept is apt.[9] Consider the following toy model of anger emotion evolution. Take Watts-Strogatz network with two phenotypes (non-emotionist P1 and emotionist/moral P2). Agents play the standard Prisoner's Dilemma with neighbours, and P1 and P2 can engage one another. Phenotypes

8. Hence, I do not commit to a particular set of emotions as *the* epistemic emotionist morality. Once functional understanding of emotions is granted, any one is in principle EGT-representable.

9. Code is available here.

differ in their payoff structures and decision rules. P1 agents receive standard PD payoffs and employ the familiar success imitation based on mean strategy payoffs in the neighbourhood. P2 agents experience emotion-adjusted payoffs: when defected against, they suffer disutility $-d_a$ and punish the opponent with probability generated by an activation function increasing in the vengeance parameter v. If punishment triggers, P2 gains additional utility v but pays cost of punishment γ, while the opponent suffers harm δ. Figure 5 contains modified payoff matrices of P1 vs P2 and P2 vs P2 encounters, with terms only realised by punishment preceded by indicator I_p (equal 1 when punishment is triggered, 0 otherwise); vanilla PD matrix would characterise P1 vs P1 encounter.

		P2	
		C	**D**
P1	**C**	$R,\ R$	$S,\ T$
	D	$T-I_p\delta,\ S-d_a+I_p(v-\gamma)$	$P,\ P-d_a$

(a) P1 vs P2

		P2	
		C	**D**
P2	**C**	$R,\ R$	$S-d_a+I_p(v-\gamma),\ T-I_p\delta$
	D	$T-I_p\delta,\ S-d_a+I_p(v-\gamma)$	$P-d_a,\ P-d_a$

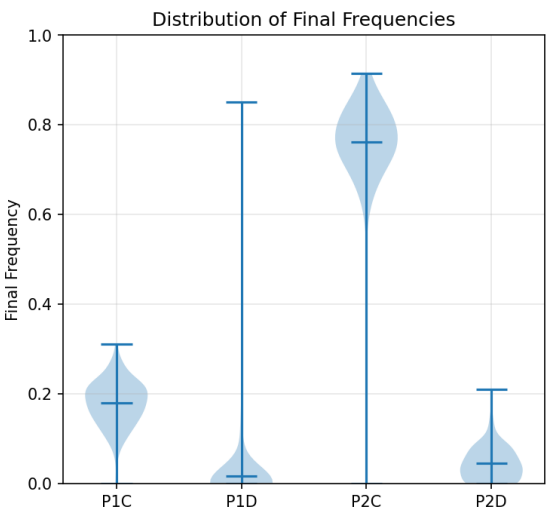
(b) P2 vs P2

Figure 5: Anger-adjusted Prisoner's Dilemma payoff matrices for both phenotypes

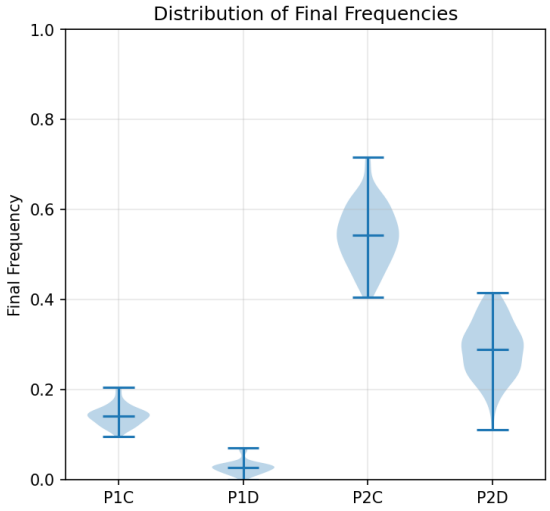
Our toy model divorces strategic and 'moral' selection. Each round, some proportion of the population updates strategies (without necessarily switching types). P1 agents adopt strategies with higher average payoff with probability proportional to the difference of said average and their latest payoff. P2 agents follow the same rule for cooperation but only adopt defection if the benefit exceeds defection aversion d_a. Further, each round, some (lower) fraction of agents is forced into changing phenotypes. This occurs as agents compare average payoffs of P1 versus P2 neighbours and switch phenotypes with probability proportional to difference between the two. This corresponds not to social learning but to natural phenotype selection in the neighbourhood.

Crucially, agents like 'non-punishing, defecting moralist' not only become a possibility but a decidedly different thing from 'defecting a-moralist', despite being behaviourally identical. Even this rough set-up enables very interesting simulation trends, like stable mixes of emotionist and non-emotionist cooperators (figure 6a), and of defecting and cooperating emotionist agents (figure 6b).[10] Interpreted against the emotionist definition of morality, the latter states are *not* completely moral as there are non-emotionist agents in them, while the former ones are completely moral, despite there being defectors in them. This marks a departure from the most plausible reading of the original models. Finally, note how unlike the classic modelling of morality, we are explicit in what 'morality' consists in for the purpose of the model-based explanation, what the desiderata for proclaiming its successful emergence are, and that we are non-arbitrary in the definition of 'morality', having adopted an existing metaethical theory.

10. On violin plots, wider sections indicate higher frequency of observations at that value.



(a) Coexistence of P1 and P2 cooperators ($v = 2, d_a = 1, \gamma = 0.5, \delta = 2$)



(b) Survival of P2 defectors ($v = 1.5, d_a = 0.1, \gamma = 0.5, \delta = 2$)

Figure 6: Plots of simulation results (200 agents, 100 repeated runs of 10000 rounds each), $T = 4,\ R = 3,\ P = 2,\ S = 1$

Let me explicate the dialectical role of this sketch. Taken as a serious attempt to model the functional description of the emotion of anger, it is woefully bad.[11] The functionality of anger is much more complex. Beyond influencing choice of action, its adequate operationalisation must at least include an embodied signalling function alerting the observers of the agent's readiness to punish future deviant behaviour. Precise character of how anger weighs on our action is also more complex: the space of parameters that the function of anger acts upon is much larger than ours, their interaction is non-linear, etc. But crucially, on the functional emotionist view, while anger is a more complicated function than the model suggests, it is still just a function. Importantly, it is a function from behaviour to behaviour, mediated by a complex (but not entirely obscure) interaction of variables bearing on the individual's decision-making. It is a more complicated thing than I have described but it is not unlike it in kind. Therefore, a more adequate model will only have to involve a more mechanically nuanced and empirically informed specification of what anger does but not a qualitatively different specification of what anger is.

Luckily, even the model sketch provided above will suffice for demonstrating how the preceding discussion allows one to dispel the pre-emptive scepticism of Alexander's argument. Even this minimal model template contains two important qualities absent from any present EGT model as used in philosophy. First, the model incorporates a mechanism that influences but does not necessarily determine the agent's action. There is now a component to the model that is always part of the relevant agent's decision-making, even if it does not, in the end, trigger the relevant action. Secondly, as a model of morality, it shifts the locus of 'morality' from the strategy profiles of the population onto the presence or absence of motivational mechanisms for strategy choice. That is, the locus of explanation is the prevalence of the emotional phenotype rather than the prevalence of moral strategies. This distinction is non-trivial, as agents of the moral phenotype may nevertheless consistently opt for immoral strategies. But given our focus on emotionist morality, what act they choose is irrelevant for whether they are moral. Instead, it is what influences their choice that decides that.

11. This is disregarding robustness and sensitivity concerns one should have about the sketch.

Can all this answer Alexander's worries that all EGT models can be interpreted in behavioural terms? After all, we seem to have done something akin to what he warned us against: we complicated the strategies slightly, relabelled some of them and introduced more psychology to the agent's decision-making mechanism. In (Alexander 2007), we are warned to not be fooled by the names of the variables like 'how good it is to punish.' At the end, these are still computational components to the agent's deterministic strategies – it is just that now, they are slightly more complicated due to the presence of additional parameters. As such, no amount of enriching the models can capture the relevant motivational structures behind the moral decision-making in the real world.

We will see this criticism lands much worse after the meta-ethical discussion we have conducted. Appreciate how much easier the assessment of adequacy and of statements like 'not capturing the motivational components' and 'really punish' becomes once we have fixed a meta-ethical view to work towards. Given the absence of a clear meta-ethical theory then, how to respond to such criticisms was unclear what how this question can be responded to now.

In light of the epistemic emotionist view and the functionalist understanding of emotions, why do real people really punish? They really punish because their decision-making is acted upon by a motivational mechanism that has an elicitor (behavioural profiles of other agents) and an action tendency (corrective other-directed action with the goal of changing conduct) (Haidt 2003). As far as the moral emotion of anger concerns behaviour, that is it – it is an input-output relation. Models such as ours capture this via a computational representation of said relation. There are agents who, in their strategy change, are causally influenced by this functional mechanism. Therefore, they exhibit an important component of morality: namely, its functional role. Note that the locus of morality in this discussion is not on behaviour but on the causal influences on behaviour. Indeed, as we saw in the provided model template, the agent can have the emotional phenotype that does not translate into the moral action but only makes it more likely, and this fact would already qualify the agent as moral.

Here, the critic might press two concerns. She may grant us that the presence of a mechanism that each agent of the moral phenotype is equipped with and that sways the agent towards moral behaviour accommodates the functional component. However, the phenomenology of emotions is nowhere to be found. So, have we really explained the evolution of emotions if we have only shown their functional component? While the critic would be correct that the phenomenology is missing, I doubt that the critic is talking about something intelligible when demanding an operationalisation of phenomenological experiences. Further, presumably, the evolution of a trait can only be explained satisfyingly if that trait influences action; as O'Connor puts it, 'when it comes to evolution, behaviour is where the rubber meets the road' (2019, p. 444). If one is interested exclusively in explaining the pure phenomenology of emotions, it is probable that an evolutionary explanation is simply not the kind of explanation being sought. The critic may continue that it sounds like our problem, since we have ourselves stated that morality is in part about phenomenological experiences. This is true and if one finds the pure phenomenology of moral emotions to be the most important constituent of morality, then EGT is of little assistance. But I think that in explaining morality, we must keep our eyes on what is puzzling about morality rather than everything there is to be said about it. And what is puzzling about morality as a set of motivating emotions is not that they feel funny but that they motivate us to do things that are often irrational and self-detrimental. What cries out for explanation is how these motivational mechanisms ever persisted in us. Covering the functional component of morality seems like covering the core of morality as a concept.

The second concern consists in applying the old criticism to new models. Computational agents in any model can be understood to both literally and figuratively be following a script.

So, psychologically enriched models are still up for interpretation in purely behavioural terms, exactly as before (Alexander 2007). This is also correct but, unlike the previous concern, in a rather uninteresting sense now. All models are idealised and yet there exist criteria for telling apart better and worse models. A crucial part of this evaluation is how well the constituents of the model map onto the constituents of the real phenomenon. Indeed, we *may* doodle on a piece of paper and call the result a model of the world economy, just like we *may* stretch our interpretation of highly detailed models of motivational structures underlying moral behaviour onto a purely behaviourist reading. Anything can be called a model of anything and yet not all models are created equally. That EGT models simply *admit* of such a behavioural interpretation, which is the extent of what Alexander suggests, is therefore not enough. The relevant question is whether the behavioural interpretation is at once the most plausible in application to models that explicitly account for the internal motivational influences on agents' actions. And the answer is no, because under the behavioural interpretation, components of the model such as the anger mechanism do not map onto anything.

Another way to see that the psychological enrichment of agential decision-making is the correct modification is to appreciate that such a model does not neatly map onto the zombie world we mentioned when explicating Alexander's objection. That is one where people behave as-if-morally but without any motivations for their actions. For instance, how to interpret in the zombie world an agent of the anger-having phenotype who nevertheless does not punish defectors due to unfavourable prospects of doing so? A purely behavioural interpretation is possible only if one ignores some components of the model completely, focusing only on what strategies are being played with no concern for what caused these strategies to be played throughout the simulation. But this seems like an invalid strategy, something very close to assuming the behavioural interpretation before consulting the model rather than the model suggesting it naturally. Instead, the most natural interpretation is one that acknowledges the strategy the agent plays in the game to have been generated through a complex decision procedure, where the precise specification of the decision procedure is suspiciously like the motivational component of the emotion of anger. That is to say, the most proximate interpretation of the model is not behavioural, but a thick one, as modelling a population of agents that do really have a complex internal motivational mechanism of anger influencing their actions.

Note that this line of response to the critic becomes stronger the more complicated the model gets. Enriching the decision-making procedure of the agent – by introducing functional dependencies on interaction history, state of the network as a whole, etc. – greatly decreases the plausibility of a purely behavioural interpretation. The latter would have to explain the choice of moral behaviour in terms of certain parameter specifications and by reference to a particular algorithm whereby these parameter values yield this behaviour. With growing complexity, we may expect this story to get so multiparameterised and involved that a mentalist, thickly moral interpretation becomes more natural and parsimonious as interpretation of the strategy choice. Indeed, the behaviourist story itself might start sounding like a description approximating that of a mental state. For an intuition boost, consider if a behaviourist interpretation strikes one as plausible when applied to ABM agents whose method of output generation is an MLP neural net (Douven 2024). In models with the relevant compelxity and realisticness of decision-making, it would be more plausible to infer that the model represents agents with the emotion of anger, not as-if anger. And the emotion of anger is part of morality. And by modelling its evolution, we can model part of morality. By repeating this strategy for all emotions that a given version of emotionism considers moral, we will have provided an EGT explanation of morality proper, as defined by the respective meta-ethical theory.

6 Conclusion

Modelling morality using EGT has been criticised for inadequate treatment of its explanandum. Further, the value of the method itself has been doubted by those who think that no plausible (i.e., non-behaviourist) conception of morality can ever be accommodated in a computational model.

The current work has shown this dire outlook to be premature. A suitably thick and independently motivated meta-ethical notion of morality can serve as the target of EGT explanations. In particular, I argued that due to the possibility of casting emotions as functions from behaviour to behaviour, epistemic emotionism can serve as such an explanandum. If EGT work focuses on more detailed and empirically nuanced representation of how emotions are elicited by and motivate behaviour, then the wide adoption of such a mechanism in a given model can serve as a faithful referent of emotionist morality. Ultimately, if one understands morality to consist in a number of select emotional mechanisms, EGT may provide reasons to think morality so-understood has evolved due to boundedly rational social learning of our ancestors.

The main ambition of this work has been conceptual: to provide a counter-example to authors who think EGT is inherently unable to model any interesting conception of morality. While I have given a proof of concept for the proposed framework, the rebuttal is incomplete without more sophisticated computational work. Additionally, modelling of any social phenomenon whose genesis can be traced to human prehistory owes an empirical validation. As I briefly mentioned, while there is much work to be done in this respect, the amount of relevant empirical evidence and scientific consensus on *some* structural characteristics of hominin societies motivate cautious optimism that empirical objections can be placated.

Finally, I hope this work has been valuable in clarifying the debate and making it easier to see what exactly the objections target in the modelling practice and what can and cannot serve as admissible answers to them.

References

Alexander, J. McKenzie. 2007. *The Structural Evolution of Morality.* Cambridge University Press.

———. 2015. "Cheap Talk, Reinforcement Learning, and the Emergence of Cooperation." *Philosophy of Science* 82, no. 5 (December): 969–982. https://doi.org/https://doi.org/10.1086/684197.

Alexander, J. McKenzie, and Brian Skyrms. 1999. "Bargaining with Neighbors: Is Justice Contagious?" *Journal of Philosophy* 96 (11): 588–598. https://doi.org/10.2307/2564625.

Arnold, Eckhart. 2008. *Explaining Altruism: A Simulation-Based Approach and its Limits.* Ontos Verlag.

Aydinonat, N. Emrah, Samuli Reijula, and Petri Ylikoski. 2021. "Argumentative Landscapes: The Function of Models in Social Epistemology." *Synthese* 199 (1-2): 369–395. https://doi.org/10.1007/s11229-020-02661-9.

Ben-Ze'ev, Aaron. 2000. *The Subtlety of Emotions.* The MIT Press. ISBN: 9780262268066.

Birch, Jonathan. 2021. "Toolmaking and the Evolution of Normative Cognition." *Biology and Philosophy* 36 (1): 1–26. https://doi.org/10.1007/s10539-020-09777-9.

Braithwaite, R. B. 1955. *Theory of Games as a Tool for the Moral Philosopher. An Inaugural Lecture Delivered in Cambridge on 2 December 1954.* Cambridge [Eng.]: University Press.

Bruner, Justin P. 2018. "Bargaining and the Dynamics of Divisional Norms." *Synthese* 197, no. 1 (February): 407–425. https://doi.org/https://doi.org/10.1007/s11229-018-1729-4.

———. 2021. "Nash, Bargaining and Evolution." *Philosophy of Science* 88, no. 5 (December): 1185–1198. https://doi.org/https://doi.org/10.1086/715778.

Clavien, Christine. 2009. "Gibbard's Expressivism: an Interdisciplinary Critical Analysis." *Philosophical Psychology* 22, no. 4 (August): 465–485. https://doi.org/https://doi.org/10.1080/09515080903153626.

D'Arms, Justin. 2000. "When Evolutionary Game Theory Explains Morality, What Does It Explain?" *Journal of Consciousness Studies* 7 (1-2): 296–299.

Douven, Igor. 2024. "Social Learning in Neural Agent-Based Models." *Philosophy of Science* 92, no. 1 (October): 1–21. https://doi.org/https://doi.org/10.1017/psa.2024.33.

Dunbar, Robin. 1993. "Coevolution of Neocortical Size, Group Size and Language in Humans." *Behavioral and Brain Sciences* 16, no. 4 (December): 681–694. https://doi.org/https://doi.org/10.1017/s0140525x00032325.

———. 2002. *Grooming, Gossip and the Evolution of Language.* Cambridge, Mass.: Harvard University Press. ISBN: 9780674363366. https://book.douban.com/subject/2282093/.

Ellemers, Naomi, Jojanneke van der Toorn, Yavor Paunov, and Thed van Leeuwen. 2019. "The Psychology of Morality: A Review and Analysis of Empirical Studies Published From 1940 Through 2017." *Personality and Social Psychology Review* 23, no. 4 (January): 332–366. https://doi.org/https://doi.org/10.1177/1088868318811759. https://journals.sagepub.com/doi/full/10.1177/1088868318811759.

Ferguson, Richard Brian. 2013a. "Pinker's List: Exaggerating Prehistoric War Mortality." *War, Peace, and Human Nature* (April): 112–131. https://doi.org/https://doi.org/10.1093/acprof:oso/9780199858996.003.0007.

———. 2013b. "The Prehistory of War and Peace in Europe and the Near East." In *War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural Views,* edited by Douglas Fry. Oxford University Press, April.

Forber, Patrick, and Rory Smead. 2014. "An Evolutionary Paradox for Prosocial Behaviour." *The Journal of Philosophy* 111 (3): 151–166. ISSN: 0022362X, accessed June 28, 2025. http://www.jstor.org/stable/43820827.

Fry, Douglas. 2006. *The Human Potential for Peace.* Oxford University Press, USA.

Gamble, Clive. 2013. *Settling the Earth: the Archaeology of Deep Human History.* Cambridge ; New York: Cambridge University Press, , Cop. ISBN: 9781107013261.

Gibbard, Allan. 1992. *Wise Choices, Apt Feelings: A Theory of Normative Judgment.* Cambridge: Cambridge Univ. Press.

Haidt, Jonathan. 2003. "The Moral Emotions." In *Handbook of Affective Sciences,* edited by R.J. Davidson, K.R. Scherer, and H.H. Goldsmith. Oxford: Oxford University Press.

Harms, William. 2000. “The Evolution of Cooperation in Hostile Environments.” *Journal of Consciousness Studies* 7 (1-2): 1–2.

Harms, William, and Brian Skyrms. 2009. “Evolution of Moral Norms.” *The Oxford Handbook of Philosophy of Biology,* 434–450. https://doi.org/10.1093/oxfordhb/9780195182057.003.0019.

Harvati, Katerina. 2007. “100 Years of Homo heidelbergensis – Life and Times of a Controversial Taxon.” https://api.semanticscholar.org/CorpusID:191055275.

Hume, David. 1896. *A Treatise of Human Nature.* Oxford: Clarendon Press, January. https://www.files.ethz.ch/isn/125487/5010_Hume_Treatise_Human_Nature.pdf.

Hutcherson, Cendri A., and James J. Gross. 2011. “The Moral Emotions: a Social–functionalist Account of Anger, Disgust, and Contempt.” *Journal of Personality and Social Psychology* 100 (4): 719–737. https://doi.org/https://doi.org/10.1037/a0022408.

Joyce, Richard. 2005. *The Evolution of Morality.* Bradford.

Kelly, Robert. 2013. *The Lifeways of Hunter-gatherers: the Foraging Spectrum.* Cambridge: Cambridge University Press. ISBN: 9781107024878.

Keltner, Dacher, and James J. Gross. 1999. “Functional Accounts of Emotions.” *Cognition & Emotion* 13, no. 5 (September): 467–480. https://doi.org/https://doi.org/10.1080/026999399379140.

Keltner, Dacher, and Jonathan Haidt. 1999. “Social Functions of Emotions at Four Levels of Analysis.” *Cognition & Emotion* 13, no. 5 (September): 505–521. https://doi.org/https://doi.org/10.1080/026999399379168.

Kirk, Robert. 1974. “Sentience and Behaviour.” *Mind* 83 (329): 43–60. ISSN: 00264423, 14602113, accessed June 27, 2025. http://www.jstor.org/stable/2252795.

Kissel, Marc, and Nam C. Kim. 2018. “The Emergence of Human Warfare: Current Perspectives.” *American Journal of Physical Anthropology* 168, no. S67 (December): 141–163. https://doi.org/https://doi.org/10.1002/ajpa.23751.

Kitcher, Philip. 1999. “Games Social Animals Play.” *Philosophy and Phenomenological Research* 59 (1): 221–228.

———. 2014. *The Ethical Project.* Harvard University Press, March. ISBN: 9780674265141.

Klein, Richard G. 2009. *The Human Career.* University of Chicago Press, April. ISBN: 9780226027524.

Levy, Arnon. 2011. “Game Theory, Indirect Modeling, and the Origin of Morality.” *The Journal of Philosophy* 108 (4): 171–187. ISSN: 0022362X, accessed February 5, 2023. http://www.jstor.org/stable/23039013.

MacDonald, Katharine, Fulco Scherjon, Eva van Veen, Krist Vaesen, and Wil Roebroeks. 2021. “Middle Pleistocene Fire Use: the First Signal of Widespread Cultural Diffusion in Human Evolution.” *Proceedings of the National Academy of Sciences* 118, no. 31 (July): e2101108118. https://doi.org/https://doi.org/10.1073/pnas.2101108118.

Maschler, Michael, Shmuel Zamir, and Eilon Solan. 2020. *Game Theory.* ISBN: 9781108636049.

O'Connor, Cailin. 2016. "The Evolution of Guilt: A Model-Based Approach." *Philosophy of Science* 83 (5): 897–908. https://doi.org/10.1086/687873.

———. 2019. "Methods, Models, and the Evolution of Moral Psychology." *arXiv (Cornell University)* (January). https://doi.org/https://doi.org/10.48550/arxiv.1909.09198.

Prinz, Jesse. 2007. *The Emotional Construction of Morals.* New York: Oxford University Press.

———. 2015. "An Empirical Case for Motivational Internalism." In *Motivational Internalism,* edited by Gunnar Björnsson, Caj Strandberg, Ragnar Francen Olinder, John Eriksson, and Frederik Björklund. Oxford University Press.

Raiffa, Howard, and R. Duncan Luce. 1957. *Games and Decisions.* New York: Wiley. ISBN: 9780471553410.

Ramsey, Grant, and Michael J. Deem. 2022. "Empathy and the Evolutionary Emergence of Guilt." *Philosophy of Science* 89 (3): 434–453. https://doi.org/10.1017/psa.2021.36.

Rodríguez, Jesús, Christian Willmes, Christian Sommer, and Ana Mateos. 2022. "Sustainable Human Population Density in Western Europe Between 560.000 and 360.000 Years Ago." *Scientific Reports* 12, no. 1 (April): 6907. https://doi.org/https://doi.org/10.1038/s41598-022-10642-w. https://www.nature.com/articles/s41598-022-10642-w.

Rosenstock, Sarita, and Cailin O'Connor. 2018. "When It's Good to Feel Bad: an Evolutionary Model of Guilt and Apology." *Frontiers in Robotics and AI* 5 (March). https://doi.org/https://doi.org/10.3389/frobt.2018.00009.

Skyrms, Brian. 1996. *Evolution of the Social Contract.* New York: Cambridge University Press.

———. 2000. "Stability and Explanatory Significance of Some Simple Evolutionary Models." *Philosophy of Science* 67, no. 1 (March): 94–113. https://doi.org/https://doi.org/10.1086/392763.

———. 2003. *The Stag Hunt and the Evolution of Social Structure.* Cambridge University Press. ISBN: 9781139165228.

Skyrms, Brian, and Robin Pemantle. 2000. "A Dynamic Model of Social Network Formation." *Proceedings of the National Academy of Sciences* 97 (16): 9340–9346. https://doi.org/10.1073/pnas.97.16.9340. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.97.16.9340. https://www.pnas.org/doi/abs/10.1073/pnas.97.16.9340.

Sterelny, Kim. 2021. *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution.* New York, Ny: Oxford University Press. ISBN: 9780197531389.

Tangney, June P., Jeffrey Stuewig, Elizabeth T. Malouf, and Kerstin Youman. 2013. "Communicative Functions of Shame and Guilt." In *Cooperation and its Evolution,* edited by Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser. MIT Press.

Thieme, Hartmut. 1997. "Lower Palaeolithic Hunting Spears From Germany." *Nature* 385, no. 6619 (February): 807–810. https://doi.org/https://doi.org/10.1038/385807a0.

Tomasello, Michael. 2016. *A Natural History of Human Morality.* Cambridge, Massachusetts ; London Harvard University Press. ISBN: 9780674088641.

Vaish, Amrisha. 2018. "The Prosocial Functions of Early Social Emotions: the Case of Guilt." *Current Opinion in Psychology* 20 (April): 25–29. https://doi.org/https://doi.org/10.1016/j.copsyc.2017.08.008.

Vanderschraaf, Peter. 2018. *Strategic Justice: Convention and Problems of Balancing Divergent Interests.* New York, NY: Oup Usa.

Weiss, Kenneth. 1984. "On the Number of Members of the Genus Homo Who Have Ever Lived, and Some Evolutionary Implications." *PubMed* 56, no. 4 (December): 637–49.

Zollman, Kevin J.S. 2008. "Explaining Fairness in Complex Environments." *Politics, Philosophy & Economics* 7, no. 1 (February): 81–97. https://doi.org/https://doi.org/10.1177/1470594x07081299.