DOI: 10.1017/psa.2026.10202
This is a manuscript accepted for publication in *Philosophy of Science*.
This version may be subject to change during the production process.

A Morality Evolutionary Game Theory Can Model

Mikhail Volkov[1]

[1]Department of Philosophy, Logic and Scientific Method, LSE
London, UK
mik_volkov@hotmail.com

Abstract

Evolutionary game-theoretic (EGT) models of morality face powerful under-addressed objections. Critics claim the simulations fail to specify their explanandum, muddying their explanatory value. Additionally, morality is suggested to be not computationally representable, jeopardising the method's general applicability. This paper explicates and addresses the objections. I argue that at least one concrete explication of morality, emotionism coupled with functionally understood emotions, can be a plausible subject of EGT explanations. I demonstrate 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.

1. Introduction

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

Attempts to explain moral behaviour despite its occasionally instrumentally counterproductive nature are not new, figuring centrally in Hume's moral philosophy, which famously supposed humans unable to satisfy their needs but through 'mutual succour'. Conventions demanding 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 abide by or break said conventions (Hume 1896, 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 evolutionary game-theoretic methods to explain 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 allowed modellers to demonstrate that in variable scenarios modelled as non-cooperative games, strategies mirroring moral behaviour can achieve wide uptake if agents follow utility-oriented rules for strategy change. These results, so goes the argument, might elucidate the origin of distributive 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 something plausibly 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 morality has since continued, the critiques have not received a convincing response showing 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 explananda and, if adequately explicated, cannot be captured by EGT. This paper responds by showing that at least one existing meta-ethical conception of morality can receive a translation into computational models. It is possible more moral theories apt for interesting EGT analyses exist. The critics' blanket scepticism, while powerful against the early EGT models, can be placated by suitably designed future models, as I show with a counter-example.

Section 2 surveys relevant EGT works and sketches their underlying explanatory structure. Section 3 outlines the criticisms against the method's explanatory value (from then on, I focus specifically on explanations of *morality*, as have the critics). After section 3, the upshot will be: EGT explanations leave their explanandum unspecified *and* we should be sceptical of the method's capacity to provide an adequate account of a phenomenon as rich as morality in general. Section 4 offers the critics a response and a positive proposal for modellers. I argue that emotionism as formulated in the meta-ethical literature coupled with a functionalist conception of emotions is an apt explication 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 EGT modelling from explanatory irrelevance (a frequent concern for the method's practitioner), and connects formal modelling work with value theory. EGT can be directly relevant to meta-ethics – but this calls for a more careful treatment of its explananda.

2. Evolutionary Game-theoretic Models of Normative Phenomena

EGT subsumes formal methods centred around boundedly rational populations of agents interacting via games. Dynamic EGT methods include variant continuous population dynamics, discrete-time models like the Moran process and agent-based local interaction models. Beyond methodological variance, EGT has been applied to distinct normative phenomena such as morality, cooperation and just social contracts based on fair bargaining. Due to underspecified explananda, whether an EGT model reflects morality or aspects of the social contract is often up to the modeller's own interpretation rather than inherent model features. Convergence to cooperation in a Prisoner's Dilemma can support the emergence of trust, a social contract or cooperation in general.

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

It is hence useful to consider EGT applications to various social phenomena, since even models not explicitly concerning morality share explanatory structure with those that do. Additionally, modellers working with thick normative phenomena of social behaviour like justice may face the same objections as will follow shortly concerning morality in particular.

Skyrms' modelling of the evolution of distributive 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 each strategy's success relative to the population average the determining factor in how widely it spreads – it is a qualitatively adaptive dynamics (Skyrms 2000). Informally, individuals performing below the population average change to strategies that outperform it. The explanatory strategy goes like this. Replicator dynamics is applied to games mirroring an interaction where some norm of justice may manifest, like the Divide-the-Cake (DtC) 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, one would standardly take the action profile (fair, fair) to constitute the just (or moral, or fair) outcome: agents split the windfall evenly.

	Demand 3 (meek)	Demand 5 (fair)	Demand 7 (greedy)
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 salience and the special attitudes towards it. Here, one can object that the salience of (fair, fair) as *the* moral outcome is culture-relative (Henrich et al. 2001) and so poses no problem for the orthodox theory. Still, insofar as each culture recognises a *particular* profile as salient, orthodox theory cannot explain *its* salience in any concrete cultural case. Whatever outcome is dictated by (even culture-relative) morality, strategic-form treatment fails to explain its salience. One hopes that embedding the game into a dynamic, social setting of EGT can yield a better explanation.

Indeed, from most initial distributions, the population ends at uniform fair sharing (figure 2). We thus have a partial explanation for the outcome's salience, at least as observed in Western morality: it is the strategy the population is likely to adopt if strategy change is qualitatively adaptive.[1] Alongside fair division, Skyrms models further demands of distributive justice, including punishments for unfair offers in the Ultimatum

[1] The worry about EGT being unable to explain the heterogeneity of 'moral' outcomes across cultures is legitimate. This is another way in which early EGT models left the explanation of morality incomplete and why any standalone EGT model might be insufficient for explaining fairness. Still, one may hold out the

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

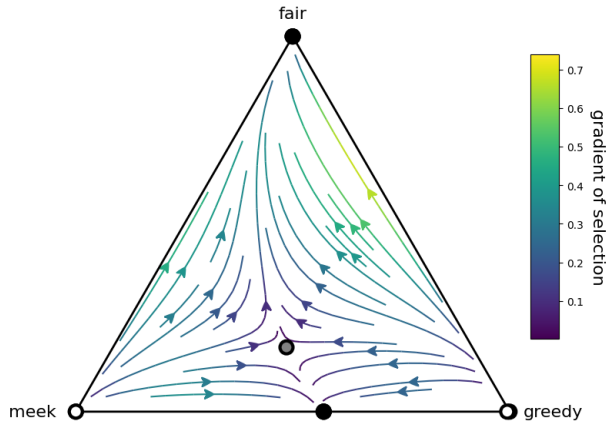


Figure 2: Simplex of Replicator Dynamics of DtC

Game and mutual cooperation in Prisoner's Dilemma that becomes sustainable under correlated interactions between proponents of the 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 better 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 dynamically changes strategies according to a success-oriented revision rule (typically imitating the most successful neighbour). Analysis proceeds by simulating the change of strategy frequencies over many runs. E.g., tracking a network playing DtC, the uniform outcome is again likely to evolve (figure 3).[2] Modifications include weighted networks of (Skyrms 2003, Part III) and (Skyrms and Pemantle 2000), where high-payoff interactions increase edge weights, raising the probability of future interactions between agents.

Most comprehensively within this strand, Alexander (2007) provides local interaction simulations across various structures and games such as Prisoner's Dilemma, Stag Hunt and others. Interestingly for us, Alexander explicitly takes *morality* for explanatory aim. So, cooperative strategies in DtC reflect the sense of *fairness*, in the Ultimatum Game *retribution*, in Stag Hunt *trust*, etc.

Recent EGT applications shy away from explananda like morality and justice.[3] Nevertheless, they retain focus on similarly complex social phenomena like just bargaining schemes or cooperation.[4] 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

hope that, once models are empirically calibrated to local, culture-specific instances of norm emergence, an ensemble of EGT models can accommodate suitably heterogenous outcomes.

[2] The simulations use Watts-Strogatz networks from (Watts and Strogatz 1998), with $p_{\text{rewiring}} = 0.5$ and connected ring as initial graph structure.

[3] Bruner (2018) puts reasons for this shift well: 'morality' as an explanandum has been tricky to pin down. This is the main issue this paper addresses.

[4] For overview of the topic see (Maschler et al. 2020, Ch. 16).

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

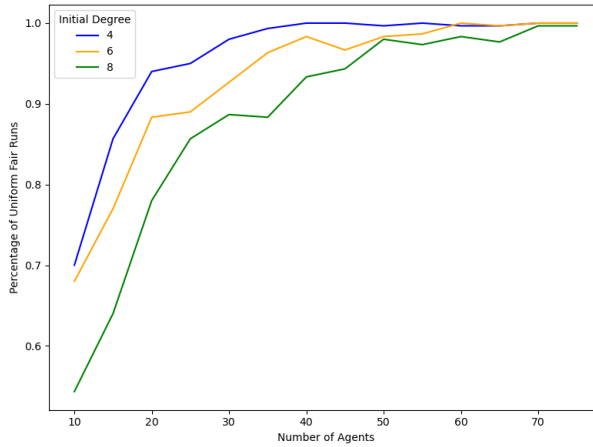


Figure 3: % of Uniform Fair Runs / Population Size for Watts-Strogatz Network Playing DtC

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?

Matthew		Luke: Play	Luke: Don't Play
	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, the Nash bargaining solution appears most likely to emerge,[5] suggesting populations gravitate to (Don't Play, Play) with 'Matthews' giving up their play time entirely.[6] Elsewhere, Bruner (2018) 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 recognised as fair outcomes, these models suggest that fairness may emerge from boundedly rational populations through repeated learning.

EGT work on bargaining has covered multiple intricate set-ups (Vanderschraaf 2018; Zollman 2008). For instance, the replicator dynamics model from (Zollman 2008) shows

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

[6] Assuming (Play, Play) as the disagreement point.

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

that fair division sometimes evolves more readily in composites of two individual games than in individual games themselves.

As subject of analysis, 'morality' can be approached in two ways: descriptive and normative (Dahl 2023; Frankena 1966). The former concerns the notion of morality as ordinarily invoked among people (i.e., what groups subjectively take to be morally significant), the latter aims to describe a collection of norms/judgements that would, under appropriate conditions, be endorsed by any moral agent (here, 'moral' is opposed to 'immoral'). The models surveyed are decidedly descriptive-elucidatory, not normative: they aim to explain observed phenomena of social interactions already delegated to the moral domain, not advance a revision of how morality ought to be conceived. In pursuing this project, EGT modellers have tended to employ different methods and focus on slightly different normative explananda. Underlying the models we just covered, however, is the following explanatory structure:

- In the real world, normative phenomenon X (morality or justice) is observed.
- In the real world, interactions transpire in repeated social settings.
- 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 through repeated interactions.[7]

EGT models based on success-oriented strategy change (which most are) will fall under a similar template. The upshot: we are fair, just and moral because we have learned, in success-guided fashion, from the population. Unfortunately, modellers generally leave implicit the connection between computational demonstrations and target phenomena (Aydinonat et al. 2021). Our case, too, allows multiple possible interpretations. I take the EGT modeller's ambition to be explaining the emergence of (whatever the modeller implicitly means by) morality and justice in a piecemeal fashion by modelling the emergence of different moral norms in varying scenarios. So, Alexander follows his piecemeal modelling saying that 'the key to our moral nature...lies in the fact that we all face repeated interpersonal decision problems – of many types – in socially structured environments' (2007, p. 278), while Harms and Skyrms seem to mean this when saying that explanation of morality must proceed in stages, 'with different explanations for different kinds of norms' (2009, p. 1). I also take the project's aim to be to explain how descriptively understood morality and justice evolved in human communities – a holistic causal explanation of real moral behaviour, with its observed and felt complexity, not pseudo-moral behaviour or alternative history for how morality could have emerged. The success of the project so conceived is at stake in the following discussion.

[7] 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).

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

3. Critique of EGT-based Explanations

EGT explanations of morality in particular have faced extensive critique divisible into two families. One focuses on empirical evaluation of the 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). Additionally, using utility-driven game-theoretic models for what is partially a fitness-based evolutionary process summons objections rooted in disanalogies between utility and fitness (Grüne-Yanoff 2011; Okasha 2016; Schulz 2014; Sugden 2001). Some existing works do contend with the empirical challenge; notably, Bowles and Gintis' book-length treatment of normative evolution (2011) focused on the context of group selection. The other family concerns the nature of the explanandum and how the models relate to it. This paper defends against the second critique, leaving the empirical challenge for future work.

The second strand contains two attacks: (1) it is unclear what 'morality' means in the modellers' works, and (2) no plausible sense of 'morality' would lend 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.

Articles supplying EGT explanations of morality, fairness and just bargaining often provide no explication or definition of their explananda. Instead, they implicitly assume explanatory success when a population in replicator dynamics or on a network widely adopts a strategy antecedently denoted as moral. But what precisely has been explained? Modellers cannot answer while 'justice' and 'morality' remain unspecified (D'Arms 2000). Consequently, assessing these models as good or bad explanations is impossible: without a clear explanatory target, there is naturally no telling whether it is reached.

D'Arms (2000) and Kitcher (1999; 2014) stress this point. While their responses concern Skyrms' work, they apply to subsequent models of morality and just bargaining also. D'Arms starts by noting that the EGT project can be interpreted in two ways: as purporting to explain the behavioural phenomenon of altruism or morality as a distinctly human phenomenon. Crucially, these come apart. Altruism refers to a pattern of behaviour, whereby agents act in favour of others at cost to themselves, and posits nothing as such about the causes of this behaviour. But description of morality is not completed by this behavioural aspect alone: morality's presence in a population must necessitate additional facts about its agents beyond behaviour. Kitcher expresses this when mentioning a 'superstructure' of normative concepts to be accounted for (1999, p. 223). These worries would survive even in *The Ethical Project*, where Kitcher reiterates the critique (§8, 9). As examples of said additional facts, critics cite punishment systems and feelings of guilt inherent in the members of the population (D'Arms 2000) and special evaluative attitudes towards other immoral agents (Kitcher 1999).

We face a fork. Interpreting the models as modelling altruism, with 'morality' and 'justice' as unfortunate shorthands for it, leaves them incomplete as representations of morality and justice, unless the latter are understood in a rather deflationary sense. After all, not all altruistic behaviour is moral (e.g., bees are presumably not moral). Models

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

of altruism could constitute a partial explanation of morality – but this explanatory relation requires considerably more fleshing out than the works surveyed do. If models are taken to demonstrate the evolution of morality as a complex of distinctively human mechanisms for sustaining concrete behavioural patterns, they are unsuccessful because models contain and explain no such mechanisms. They contain no complex of things that one can recognise as referents of the term 'morality' and so fail to explain it.

D'Arms notes a further complication: models disregarding mechanisms supporting real-world morality constitute a modus tollens against any explanation based on them. This is because we *know* that real moral norms subsist on much besides prudential learning and behavioural changes. In contrast, the 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', collapsing morality into expedience (D'Arms 2000, p. 298). Thus conceived, EGT not only leaves the thickness of phenomena under discussion unexplained but suggests the redundancy of the motivational patterns supporting real moral sense.

The upshot: if modellers are agnostic about the meaning of their own explanandum, assessing their models (and what they are models *of*) becomes impossible. In turn, insofar the intended interpretation concerns distinctively human moral behaviour, the models fail to explain it because they misrepresent its central aspects.

3.2.

Alexander's extension of the objection (2007, Ch. 8) goes further and deserves closer scrutiny. While D'Arms only targets Skyrms' early models and is overall optimistic about the promise of EGT, Alexander thinks the critique covers EGT modelling of morality as such. The scepticism stems from an alleged inherent mismatch between method and explanandum. All suitable notions of morality, goes the argument, are non-behavioural and contain motivational or superstructual components. For instance, we do not just punish and proceed about our day; we *want* to punish, enjoy seeing wrongdoers receive just deserts and feel strongly that punishment must occur. 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 and crucial to many meta-ethical views.

Alexander insists EGT cannot adequately incorporate these superstructural components, being fundamentally behaviour-centred: payoffs accrue to strategies in the stage game and are the sole factor in strategy change. 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 concludes that models will inevitably underdetermine 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*. (Alexander 2007, p. 273)

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

This appeal needs precisification. The point about the models invariably admitting behavioural interpretation may be illustrated by an example familiar from philosophy of mind (Kirk 1974). Imagine a zombie world where humans behave as we do in morally relevant cases but lack motivations underlying their behaviour. Perhaps as the 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 recipe maximises their success under constraints of social structure and bounded rationality, then start their day. These people are expedient rule-followers, not moral. There is no morality in a world like this, despite the inhabitants' behaviour matching that of moral agents. However, EGT models apply equally well to the zombie world; observed behaviour there is as well explained as that in our world. More generally, as-if-moral behaviour is multi-realisable: it can stem from a number of mechanisms, some not necessarily moral. Focusing exclusively on behaviour, 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, Alexander claims that no thick notion of morality can be adequately represented by an EGT model.

The criticisms connect to debates surrounding motivational internalism (Mackie 1977; Smith 1995; Shafer-Landau 2000). Concretely, one must be somewhat sympathetic to motivational internalism to worry about equal applicability to the zombie world or, more generally, admit conceptual mismatch between morality properly understood and whatever existing EGT models manage to showcase. If one's meta-ethical commitments imply no connection between moral sense and motivation, Alexander and D'Arms' critique appears less troubling. Given our aim to defend EGT foundationally, I go along with the critics, recognising the absence of internal motivation as a genuine problem, seeing especially how this position is far from marginal in meta-ethics. Additionally, while externalists deem moral judgements without motivation possible, they do not necessarily deny moral motivation's existence and importance to morality (Prinz and Nichols 2010).

4. A Morality EGT Can Model

The criticisms certainly paint dark prospects: *current* EGT models lack a clear explanandum and hence do not explain morality, nor should we 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 game-theoretic tradition of explaining normative phenomena, few systematic responses have appeared, perhaps due to some remaining ambiguity in the criticisms. For instance, what would it mean to fix a conception of morality? What conceptions can the modeller 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? Models allegedly

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

overlook some necessary condition for there being real morality but nothing nearing a precise indication of what has been overlooked is given, beyond broad gestures 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.

I propose a strategy responding to both concerns on their most plausible interpretations. I isolate a concrete, independently motivated notion of morality that accounts for our intuitions about its thick character (thus accommodating objections in 3.1) and argue that 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

What may the modeller offer regarding D'Arms' complaint about lack of specificity? Clearly, some notion of morality that does not deflate the term into 'a group of agents *behaving* a certain way'. It further has to provide a cluster of concepts with which to bind the term 'morality' for the purposes of a model-based explanation – that is, some meta-ethical commitment to what morality is.

If not in its behavioural dimensions, where else to locate the referent of the word 'morality'? Or: given a population of real agents, what would it take to conclude that these agents are not merely expedient but moral in the thick sense that we attribute to humans?

Numerous authors in philosophy and psychology have attempted to circumscribe uniting features of different moralities descriptively understood. Since what individuals and groups consider morally significant differs substantially, definitions aim to be high-level to accommodate this variation. For instance, Dahl (2023) defines morality as 'obligatory concerns with others' welfare, rights, fairness, and justice, as well as the reasoning, judgement, emotions, and actions that spring from those concerns'. Dahl's definition is substantive, presupposing clarity about what notions like 'justice' and 'rights' are, and builds upon Turiel's influential definition of morality as prescriptive judgements about justice, rights and welfare (though without invoking emotions or actions) (Turiel 1983). Other definitions proceed from the hypothesised function of morality. Haidt (2008) delegates to it all values, practices and psychological mechanisms suppressing free-riding and selfish concerns, whereby, in his view, one of these mechanisms are evaluative feelings that we take for moral judgement (Haidt and Bjorklund 2008). Some definitions articulated over the years, like one by Curry (2016) proclaiming morality to be solutions to cooperative problems, read very closely to our game-theoretic starting point (and may be subject to the same concerns).

However, scepticism looms over whether neat delineation of moral domain can be had to begin with. Some argue it cannot (Stich 2018; Machery 2018): judgements, situations and properties classified as 'moral' across cultures may simply share no underlying unifying features, such as 'being about justice and rights' or 'mechanisms enabling cooperation'. Given the cross-cultural and even inter-personal variability in what is and is not of moral significance, no clean cleaving off of moral domain may be possible. Separating it from the normative-but-not-moral domain is particularly problematic.

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

Perhaps we can only hope for a definition with blind spots that enables interesting analysis. Our paper is premised upon this hope: I will isolate one possible disambiguation of what morality is about among many. I do not claim it to be definitively convincing for meta-ethical debates; what I do claim is that it is (a) not implausible and (b) EGT-representable. This move does not seem too outrageous, '[t]here is no shame in settling for an inquiry-specific… definition' (Dahl 2023, p. 56). To carry this out, I begin with EGT critics' own implications on what original models missed, since, after all, it is these critics that my explication must placate.

From the preceding discussion in EGT literature, one inherits two other potential dimensions that 'morality' concerns as a term: the phenomenological and the functionalist (recall the critics' talk of 'motivational structures' (Alexander 2007, p. 273) and 'superstructures' (Kitcher 1999)). Can a computationally representable concept of morality be rooted in either?

First, morally significant situations in the real world are accompanied by internal feelings: satisfaction when an immoral person gets punished, guilt when we violate a norm. If such feelings are absent, then perhaps an important referent of 'morality' does not obtain and the judgement/act/situation at hand are best described as belonging to a less grave normative domain than the moral one.

Secondly, morality has a functional or dispositional role distinct from both behavioural manifestation and phenomenology. That is, if a person is moral, they must, in situations of a certain structure, feel compelled or motivated to act in certain ways. Importantly, this does not necessarily translate to behaviour: 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, a part of the equation. Agents with Western morality in a Divide-the-Cake situation might not go for the fair split – but will likely feel some pull towards it or discomfort when looking back on an unfair choice. Furthermore, the presence of feelings cannot substitute this functional role. The quale of guilt – 'what-it-is-like-to-feel-guilty' – has no intrinsic relation to action. We may think otherwise because guilt is almost inevitably accompanied by morality's functional push kicking in. When I feel guilty, I am simultaneously more likely to apologise; I really want to do so. But it is sensible to divorce the feeling itself from the disposition to act that accompanies it. The separate functional role of morality is precisely this being compelled 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 these two extant components of morality that are arguably more important than the behavioural aspect for diagnosing morality in a population. Indeed, purely behavioural data may mislead about moral status. Consider citizens forced by a dictatorial regime into inhumane activities like snitching on their neighbours, despite feeling immense guilt and desire to stop. Despite no behavioural manifestation being present, there *is* morality in this population – classifiable as an instance of akrasia, an unambiguously moral phenomenon.

Consequently, modellers are well-advised to focus on capturing the other dimensions of morality. However, computational models cannot in principle track the emergence of phenomenological experiences. Indeed, it is unclear what this would even mean. One 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. However, D'Arms asks for a concrete meta-ethical

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

commitment – a clear disambiguation of what morality is. The next subsection will outline such a conception rooted in the motivational dimension of morality. If such a theory can be found and translated into a model, then there may exist EGT models to which one can 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 *M*-morality, not as-if-morality.'

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 clarifying what our EGT models explain. In turn, if we want a meta-ethical theory that ties morality to motivational structures, *emotionist* theories are natural candidates. Indeed, my claim, which I shall justify shortly, is that they are suitable explananda for EGT models.

The choice of emotions-centred theory is salient given 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 economic analysis of emotions as mechanisms for cementing cooperative outcomes (Frank 1988; Hirshleifer 1984). Finally, our choice is suggested by recent EGT literature, where the evolution of moral psychology has come into focus (O'Connor 2019, 2016; Cimpeanu et al. 2025).

The label of emotionism subsumes many meta-ethical views, with Prinz' and Gibbard's work as perhaps most famous (Prinz 2007, 2015; Gibbard 1992). However, many authors of differing meta-ethical colours admit the crucial role of emotions for a full-bodied concept of morality (e.g., Joyce 2005). Such authors align with the broad meta-ethical position that Prinz dubs epistemic emotionism and defines so (Prinz 2007, p. 16):

Epistemic Emotionism: Moral concepts are essentially tied to emotions.

Epistemic emotionism (so called because it characterises how one *grasps* notions like 'good' and 'bad') can be further precisified, depending on how 'essentially tied' is cashed out. A naive reading is constitutive: properties, judgements and acts making up the domain of morality are those accompanied (in the case of acts) or constituted by emotional states. So, the judgement 'genocide is bad' is a negatively valenced emotional state evoked by an act of genocide, which itself is descriptively moral in virtue of evoking these emotions. Expressivists accept epistemic emotionism in their analysis of moral judgements, though some more nuancedly. For Gibbard, moral judgements express acceptance of norms about appropriateness of certain emotions. An alternative, sentimentalist reading is dispositionalist (indeed, sentimentalists define 'sentiment' as 'emotional disposition' (Prinz 2007, p. 97)): a notion falls under the domain of morality if emotions *tend* to accompany it.

In what follows, I most plausibly succeed in showing that EGT can model emotionist morality understood via less complex readings, like the naive or sentimentalist

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

versions.[8] Such theories have a drawback: there are immediate cases where they misfire as definitions of morality. Thinking back on a moral wrongdoing from twenty years ago, we may experience no emotions. Alternatively, tripping ungracefully before a romantic crush might make us intensely ashamed but without rendering the situation moral.

As mentioned, any definition of morality may be doomed to such blind spots. However, since explication is attempted, it is only proper I briefly show that emotionism is not obviously implausible as one. Turn to Carnap's criteria for explication (1950). As an explication of morality, emotionism is *fruitful*, both for the purpose at hand (as we will see shortly, it lends itself to computational analysis) and generally, as it allows for engaging future and existing psychological research in the analysis of morality. It is importantly more fruitful as subject of analysis than unspecified morality. This is because emotionism is more *exact* than the explicandum (morality-in-vacuum). It puts its neck forward as to what exactly the phenomenon in question is. Is emotionism *similar* to the original notion? Bracketing the blind spots covered, moral domain is at least partially characterised by special attitudes that tend to be more intense and action-conducive than attitudes towards routine facts like 'the pen is on the table' or those related to other normative domains, like etiquette. Indirect evidence for the centrality of emotions and motivations for the folk notion of morality is given by the myriad of theories stemming from Hume, Hutcheson and other Scottish Enlightenment figures based on it, and by motivational internalism remaining a widespread meta-ethical commitment even today. Finally, while properly evaluating *simplicity* requires a comparative analysis, emotionism, especially in its naive and sentimentalist reading, should strike one as a simple definition: moral judgements, acts and properties are those that tend to be accompanied or constituted by presence of emotions we antecedently designate as moral; morality is the collection of such judgements, acts and properties. If anything, the weakness of such an explication is excessive simplicity and insufficient exactness.

For all advantages and disadvantages of emotionism, my argument does not, strictly speaking, need the theory of morality that EGT can model to be good. Emotionism serves a counterexemplary purpose, insofar as the sceptics profess doubt on whether *any* philosophically interesting understanding of morality can be modelled. The current work attempts to prove sceptics wrong, not emotionists right. One may legitimately wonder whether vindicating EGT explanations is at all useful if the accommodated theory is independently implausible. Fortunately, although not the dominant meta-ethical stance, emotionism is not implausible in this way. Furthermore, EGT models of emotionist morality may well be co-opted by proponents of other definitions also, seeing how many meta-ethical theories incorporate emotions as important constituents. Indeed, some non-emotionist definitions – e.g., (Dahl 2023) who defines morality as about 'concerns' – seem related to versions of emotionism that understand concerns to be represented by emotions (Prinz 2007, p. 63). EGT's explanatory promise for meta-ethics does not stand and fall with the plausibility of emotionism. Thus, although the current argument will not convince everyone that the original EGT project can succeed in explaining *a very plausible* notion of morality, it will hopefully demonstrate the possibility of EGT to assist explaining *a* morality. Still, the plausibility of an explanandum is a secondary matter for our argument; the existence of one is primary.

[8] This is as opposed to theories like Gibbard's, on which, strictly speaking, morality is not exactly constituted *by* emotions but necessitates beliefs *about* emotions.

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

Now, why think that moral emotions are translatable into EGT models? Emotions are things we feel – and the locus of morality as we identified it lies in motivation. Here, the functionalist view of emotions enters. Let us take the two emotions that Gibbard considers crucial: anger and guilt (Gibbard 1992; Clavien 2009). These picks are apt given moral emotions are generally split into reflexive (self-directed) and reactive (other-directed) categories (e.g., Prinz 2007; Ben-Ze'ev 2000; Ellemers et al. 2019), of which guilt and anger are two respective instances.

The prevalent trend in the psychological literature on emotions is to define them as functions from input to behavioural output. For instance, Haidt's influential classification specifies each emotion by an elicitor and an action tendency (2003). Ramsey and Deem (2022) cast guilt as a function that takes one's wrongdoing as 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. The same holds for anger or indignation. Its output concerns incentivisation of active interaction to correct others' 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. Nothing too mystical characterises the behavioural workings of emotions in the psychological literature. All operate with either structural ('goal blockage') or behavioural terms ('threatening action'). This functionalist reading of emotions would seem to lend itself to translation into computational vocabulary.

Hence, emotionism can serve as a target for EGT if it is seeking a concept of morality to explain properly, not as-if. First, emotionism disambiguates the expected functional mechanism in moral populations in a computationally representable way. If emotions are understood functionally as they are in the science under whose domain the analysis of emotions falls, they are functions from behaviour to behaviour. This is a representation of emotions that is encodable into a computational model. Second, this concept of morality breaks 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, these explanations can be joined into what, on the most straightforward emotionist picture, amounts to an explanation of the fundamental moral mechanisms.[9] This moves EGT closer to delivering on its claim to relevance for explaining morality.

5. Back to the Pre-emptive Challenge

We 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 associating morality with some strategy frequency.

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

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

While the present paper is mainly conceptual and programmatic, a proof of concept is apt.[10] Consider this toy model of anger evolution. Take a Watts-Strogatz network (initial degree 6, $p_{\text{rewiring}} = 0.1$) with two phenotypes: non-emotionist P1 and emotionist/moral P2. Agents play standard Prisoner's Dilemma with neighbours, and P1 and P2 can engage one another. Phenotypes differ in payoff structures and decision rules. P1 agents receive standard PD payoffs and employ the familiar success imitation based on neighbourhood mean strategy payoffs. 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 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 to 1 when punishment is triggered, 0 otherwise); vanilla PD matrix would characterise P1 vs P1 encounter.

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

(a) P1 vs P2

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

(b) P2 vs P2

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

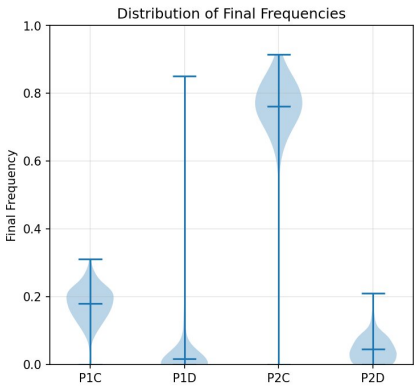
Our toy model divorces strategic and 'moral' selection. Each round, some proportion of the population updates strategies (without necessarily switching types). P1 agents adopt strategies with higher average payoff with probability proportional to the difference of said average and their latest payoff. 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

[10] Code is available at: https://github.com/Horaso-2/egt_morality

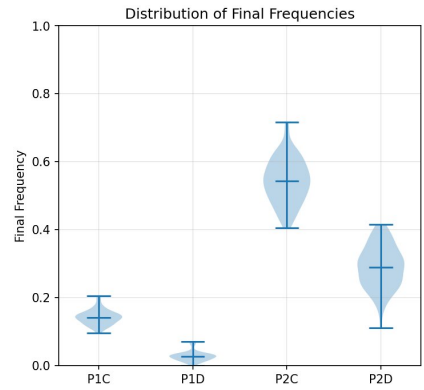
https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

switch phenotypes with probability proportional to their difference. This corresponds to natural phenotype selection in the neighbourhood, not social learning.

Crucially, agents like 'non-punishing, defecting moralist' become not only possible but a decidedly different thing from 'defecting a-moralist', despite behavioural identity in the stage game. Even this rough set-up enables interesting trends, like stable mixes of emotionist and non-emotionist cooperators (figure 6a), and of defecting and cooperating emotionist agents (figure 6b).[11] Interpreted against the emotionist definition of morality, the former states are *not* emotionist-moral as there are non-emotionist agents in them, while the latter ones are completely moral, despite there being defectors in them. This marks a departure from the original models. Unlike old models, ours is also explicit about what 'morality' consists of for the purpose of the model-based explanation and what desiderata govern its successful emergence. Finally, our definition of 'morality' is non-arbitrary, as we have adopted an existing meta-ethical theory for it.



(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 clarify this sketch's dialectical role. As a serious attempt to model anger's functional description, it is woefully bad.[12] The functionality of anger is much more complex. Its adequate operationalisation must at least additionally include an embodied signalling function alerting the observers of the agent's readiness to punish future deviant behaviour. How anger weighs on action is also more complex: the parameter space that the function of anger acts upon is much larger, parameter interactions are non-linear, etc. But crucially, on the functional emotionist view, anger is still just a function, albeit more complicated. 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 is not unlike it in kind. Therefore, a more adequate model will have to involve a more

[11] On violin plots, wider sections indicate higher frequency of observations at that value.

[12] This is disregarding robustness and sensitivity concerns one should have about the sketch.

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

mechanically nuanced and empirically informed specification of what anger does but not a qualitatively different specification of what anger is.

Fortunately, even this sketch suffices to demonstrate how the preceding discussion enables one to dispel Alexander's pre-emptive scepticism. Our minimal model contains two important qualities absent from the surveyed EGT models. First, the model incorporates a mechanism that influences but does not necessarily determine the agent's action. There is now a model component that is always part of the relevant agent's decision-making, even if it does not reliably trigger the relevant action. Second, it shifts the locus of 'morality' from the population's strategy profiles in the stage game 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 actions. 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 distinct from whether they are moral. Instead, it is what influences their choice that determines moral status.

Can this answer Alexander's worries about all EGT models admitting behavioural interpretations? After all, we seem to have done what he warned us against: complicating the strategies, relabelling some of them and introducing more psychology to the agent's decision-making mechanism. In (Alexander 2007), one is warned against being fooled by variable names like 'how good it is to punish'. Ultimately, these are still computational components to the agent's deterministic strategies – it is just that now, they are complicated by additional parameters. As such, no amount of enriching the models can capture the relevant motivational structures behind real-world moral decision-making.

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. Without a clear meta-ethical theory then, how to respond to such criticisms was unclear; with one, a response becomes possible.

On epistemic emotionism with the functionally understood emotions, why do real people really punish? They really punish because their decision-making is acted upon by a motivational mechanism with 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 like ours capture this via a computational representation of that relation. Agents causally influenced by this functional mechanism exhibit an important component of morality: namely, its functional role. Note that the locus of morality here is not behaviour but causal influences on it. As we saw in the model, the emotional phenotype does not necessarily translate into the moral action but only makes it more likely. The possession of this phenotype would already qualify the agent as emotionist-moral.

The critic might press two concerns. She may grant that a mechanism swaying 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 phenomenology is indeed missing, I doubt that the critic is talking about something intelligible when demanding an operationalisation of phenomenological experiences. Further, presumably, evolution of a trait can only be explained satisfyingly if that trait

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

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 what is sought. The critic may continue that it sounds like our problem, since we ourselves have stated that morality partly concerns phenomenological experiences. This is true and if one finds the pure phenomenology of moral emotions to be morality's most important constituent, then EGT is of little assistance, this has to be conceded. But explaining morality requires focusing on what is puzzling about it 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 often motivate irrational and self-detrimental actions. What cries out for explanation is how these motivational mechanisms ever persisted in us. Covering the functional component seems like covering the conceptual core of morality, while explaining its phenomenology does not.

The second concern applies the old criticism to new models. Computational agents in any model can be understood to both literally and figuratively follow a script. So, psychologically enriched models are interpretable in purely behavioural terms, exactly as before (Alexander 2007). This is correct but, unlike the previous concern, only uninterestingly so. A crucial part of telling good models from bad ones is how well their parts map onto the real phenomenon. Anything can be called a model of anything but not all models are created equally. That EGT models simply *admit* a behavioural interpretation, which is the extent of what Alexander suggests, is insufficient. The question is whether the behavioural interpretation is the most plausible for models that explicitly account for the internal motivational influences on agents' actions; whether it is the application these models actively suggest. The answer is no, because under the behavioural interpretation, components of the model like the anger mechanism map onto nothing. Another way to see the benefit of the psychological enrichment of agential decision-making: such models do not neatly map onto the zombie world. 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 more sophisticated version of this criticism could argue an amoral psychopath might perfectly well learn to mimic the decision-making procedure of the 'emotionist' type, so there is no problem with reading a P2 cooperator as an expedient psychopath rather than an emotionally compelled moralist. Given the computational nature of our and any future models, some form of this criticism can always be pushed. As indicated, it would be naive to claim that my model manages to capture the full functional complexity of anger. Still, no conceptual impossibility prevents increasing the empirical accuracy and complexity in operationalising the social function of anger to where a suggestion that it can be expediently mimicked becomes too far of a stretch.[13] Note that this response strengthens as models become more complicated. 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

[13] An anonymous reviewer's point suggests driving the operationalisation of strategies closer to reality through incorporating behavioural differences between moral and psychopathic individuals. An aspect of amoral strategy worthy of attention here concerns the mechanisms of learning to mimic the genuinely moral strategy. Such future work may be useful in explaining the persistence of both types as observed in actual societies.

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

behaviour in terms of certain parameter specifications and by reference to a particular algorithm whereby these parameter values yield this behaviour. With growing complexity, this story may grow 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 approximating a mental state description. Insisting on the behavioural reading indefinitely would mean reading it into the model rather than the model suggesting it naturally. For an intuition boost, consider if a behaviourist interpretation strikes one as plausible when applied to agents whose method of output generation is a multi-layered perceptron (MLP) neural net (Douven 2024). In sufficiently complex, realistic models, 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; modelling its evolution is modelling part of morality. Repeating this strategy for all emotions that a given version of emotionism considers moral provides an EGT explanation of morality as defined by at least one meta-ethical theory returning thickness to the notion.

6. Conclusion

EGT modelling of morality has been criticised for inadequate treatment of its explananda. 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.

This paper has shown such dire outlook is premature. EGT can aim to explain at least one thick meta-ethical definition of morality. In particular, casting emotions as behaviour-to-behaviour functions, versions of emotionist theories can serve as such an explanandum. If EGT work focuses on empirically validated representations of how emotions are elicited by and motivate behaviour, wide adoption of such a mechanism in a given model will faithfully reflect 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 evolved due to boundedly rational social learning of our ancestors.

Our main ambition was conceptual: to provide a counter-example to claims that EGT is inherently unable to model any interesting conception of morality. While I have given a proof of concept, the rebuttal remains incomplete without more sophisticated computational work. To fully deliver on the explanatory promise, meta-ethical conceptions other than naive emotionism would have to be analysed for EGT-representability. Additionally, modelling of any social phenomenon traceable to human prehistory owes an empirical validation.

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

References

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

Alexander, J. M. (2015) Cheap talk, reinforcement learning, and the emergence of cooperation. *Philosophy of Science*. *82*(5), 969–982. https://doi.org/10.1086/684197.

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press

Alexander, J. M. and Skyrms, B. (1999) Bargaining with neighbors: Is justice contagious? *Journal of Philosophy*. *96*(11), 588–598. 10.2307/2564625.

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

Aydinonat, N. E., Reijula, S. and Ylikoski, P. (2021) Argumentative landscapes: The function of models in social epistemology. *Synthese*. *199*(1-2), 369–395. 10.1007/s11229-020-02661-9.

Ben-Ze'ev, A. (2000) *The Subtlety of Emotions*. The MIT Press.

Bowles, S. and Gintis, H. (2011) *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press. Princeton ; Oxford.

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

Bruner, J. P. (2018) Bargaining and the dynamics of divisional norms. *Synthese*. *197*(1), 407–425. https://doi.org/10.1007/s11229-018-1729-4.

Bruner, J. P. (2021) Nash, bargaining and evolution. *Philosophy of Science*. *88*(5), 1185–1198. https://doi.org/10.1086/715778.

Carnap, R. (1950) *Logical Foundations of Probability*. Chicago University of Chicago Press.

Cimpeanu, T., Pereira, L. M. and Han, T. A. (2025) The evolutionary advantage of guilt: Co-evolution of social and non-social guilt in structured populations. *Journal of The Royal Society Interface*. *22*(228). https://doi.org/10.1098/rsif.2025.0164.

Clavien, C. (2009) Gibbard's expressivism: an interdisciplinary critical analysis. *Philosophical Psychology*. *22*(4), 465–485. https://doi.org/10.1080/09515080903153626.

Curry, O. (2016) Morality as cooperation: A problem-centred approach. pp. 27–51. 10.1007/978-3-319-19671-8˙2.

Dahl, A. (2023) What we do when we define morality (and why we need to do it). *Psychological Inquiry*. *34*(2), 53–79. https://doi.org/10.1080/1047840x.2023.2248854.

D'Arms, J. (2000) When evolutionary game theory explains morality, what does it explain? *Journal of Consciousness Studies*. *7*(1-2), 296–299.

Douven, I. (2024) Social learning in neural agent-based models. *Philosophy of Science*. *92*(1), 1–21. https://doi.org/10.1017/psa.2024.33.

Ellemers, N., van der Toorn, J., Paunov, Y. and van Leeuwen, T. (2019) The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*. *23*(4), 332–366. https://doi.org/10.1177/1088868318811759.

Forber, P. and Smead, R. (2014) An evolutionary paradox for prosocial behaviour. *The Journal of Philosophy*. *111*(3), 151–166.

Frank, R. H. (1988) *Passions Within Reason: the Strategic Role of the Emotions*. Norton.

Frankena, W. K. (1966) The concept of morality. *Journal of Philosophy*. *63*(21), 688–696. 10.2307/2024163.

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

Grüne-Yanoff, T. (2011) Evolutionary game theory, interpersonal comparisons and natural selection: A dilemma. *Biology and Philosophy*. *26*(5), 637–654. https://doi.org/10.1007/s10539-011-9273-3.

Haidt, J. (2003) The moral emotions In *Handbook of Affective Sciences*, Davidson, R., Scherer, K., and Goldsmith, H. (eds). Oxford University Press. Oxford.

Haidt, J. (2008) Morality. *Perspectives on Psychological Science*. *3*(1), 65–72. https://doi.org/10.1111/j.1745-6916.2008.00063.x.

Haidt, J. and Bjorklund, F. (2008) Social intuitionists answer six questions about morality In *Moral Psychology Vol. 2*, Sinnott-Armstrong, W. (eds). MIT Press.

Harms, W. (2000) The evolution of cooperation in hostile environments. *Journal of Consciousness Studies*. *7*(1-2), 1–2.

Harms, W. and Skyrms, B. (2009) Evolution of moral norms. *The Oxford Handbook of Philosophy of Biology*. p. 434–450. 10.1093/oxfordhb/9780195182057.003.0019.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. and McElreath, R. (2001) In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*. *91*(2), 73–78. https://doi.org/10.1257/aer.91.2.73.

Hirshleifer, J. (1984) On the emotions as guarantors of threats and promises. UCLA Economics Working Papers 337. UCLA Department of Economics.

Hume, D. (1896) *A Treatise of Human Nature*. Oxford: Clarendon Press.
Hutcherson, C. A. and Gross, J. J. (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/10.1037/a0022408.
Joyce, R. (2005) *The Evolution of Morality*. Bradford.
Keltner, D. and Gross, J. J. (1999) Functional accounts of emotions. *Cognition & Emotion*. *13*(5), 467–480. https://doi.org/10.1080/026999399379140.
Keltner, D. and Haidt, J. (1999) Social functions of emotions at four levels of analysis. *Cognition & Emotion*. *13*(5), 505–521. https://doi.org/10.1080/026999399379168.
Kirk, R. (1974) Sentience and behaviour. *Mind*. *83*(329), 43–60.
Kitcher, P. (1999) Games social animals play. *Philosophy and Phenomenological Research*. *59*(1), 221–228.
Kitcher, P. (2014) *The Ethical Project*. Harvard University Press.
Levy, A. (2011) Game theory, indirect modeling, and the origin of morality. *The Journal of Philosophy*. *108*(4), 171–187.
Machery, E. (2018) Morality: A historical invention In *Atlas of Moral Psychology*, Gray, K. and Graham, J. (eds). The Guilford Press. p. 259–265.
Mackie, J. L. (1977) *Ethics: Inventing Right and Wrong*. Penguin. Harmondsworth.
Maschler, M., Zamir, S. and Solan, E. (2020) *Game Theory*.
O'Connor, C. (2016) The evolution of guilt: A model-based approach. *Philosophy of Science*. *83*(5), 897–908. 10.1086/687873.
O'Connor, C. (2019) Methods, models, and the evolution of moral psychology. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1909.09198.
Okasha, S. (2016) Biology and the theory of rationality In *How Biology Shapes Philosophy*, Livingstone Smith, D. (eds). Cambridge University Press.
Prinz, J. (2007) *The Emotional Construction of Morals*. Oxford University Press. New York.
Prinz, J. (2015) An empirical case for motivational internalism In *Motivational Internalism*, Björnsson, G., Strandberg, C., Olinder, R. F., Eriksson, J., and Björklund, F. (eds). Oxford University Press.
Prinz, J. and Nichols, S. (2010) Moral emotions. *The Moral Psychology Handbook*. p. 111–146. https://doi.org/10.1093/acprof:oso/9780199582143.003.0005.
Raiffa, H. and Luce, R. D. (1957) *Games and Decisions*. Wiley. New York.
Ramsey, G. and Deem, M. J. (2022) Empathy and the evolutionary emergence of guilt. *Philosophy of Science*. *89*(3), 434–453. 10.1017/psa.2021.36.
Rosenstock, S. and O'Connor, C. (2018) When it's good to feel bad: an evolutionary model of guilt and apology. *Frontiers in Robotics and AI*. *5*. https://doi.org/10.3389/frobt.2018.00009.
Schulz, A. W. (2014) Niche construction, adaptive preferences, and the differences between fitness and utility. *Biology and Philosophy*. *29*(3), 315–335. 10.1007/s10539-014-9439-x.
Shafer-Landau, R. (2000) A defense of motivational externalism. *Philosophical Studies*. *97*(3), 267–291. 10.1023/a:1018609130376.
Skyrms, B. (1996) *Evolution of the Social Contract*. Cambridge University Press. New York.
Skyrms, B. (2000) Stability and explanatory significance of some simple evolutionary models. *Philosophy of Science*. *67*(1), 94–113. https://doi.org/10.1086/392763.
Skyrms, B. (2003) *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.
Skyrms, B. and Pemantle, R. (2000) A dynamic model of social network formation. *Proceedings of the National Academy of Sciences*. *97*(16), 9340–9346. 10.1073/pnas.97.16.9340.
Smith, M. (1995) Internal reasons. *Philosophy and Phenomenological Research*. *55*(1), 109–131. 10.2307/2108311.
Stich, S. (2018) The quest for the boundaries of morality In *The Routledge Handbook of Moral Epistemology*, Zimmerman, A., Jones, K., and Timmons, M. (eds). Routledge.
Sugden, R. (2001) The evolutionary turn in game theory. *Journal of Economic Methodology*. *8*(1), 113–130.
Tangney, J. P., Stuewig, J., Malouf, E. T. and Youman, K. (2013) Communicative functions of shame and guilt In *Cooperation and its Evolution*, Sterelny, K., Joyce, R., Calcott, B., and Fraser, B. (eds). MIT Press.
Turiel, E. (1983) *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.
Vaish, A. (2018) The prosocial functions of early social emotions: the case of guilt. *Current Opinion in*

Psychology. *20*, 25–29. https://doi.org/10.1016/j.copsyc.2017.08.008.
Vanderschraaf, P. (2018) *Strategic Justice: Convention and Problems of Balancing Divergent Interests*. Oup Usa. New York, NY.
Watts, D. J. and Strogatz, S. H. (1998) Collective dynamics of "small-world" networks. *Nature*. *393*(6684), 440–442.
Zollman, K. J. (2008) Explaining fairness in complex environments. *Politics, Philosophy & Economics*. *7*(1), 81–97. https://doi.org/10.1177/1470594x07081299.

https://doi.org/10.1017/psa.2026.10202 Published online by Cambridge University Press