Building machines that learn and think about morality

Christopher Burr [1] and **Geoff Keeling** [2]

Abstract. Lake et al. [30] propose three criteria which, they argue, will bring artificial intelligence (AI) systems closer to human cognitive abilities. In this paper, we explore the application of these criteria to a particular domain of human cognition: our capacity for moral reasoning. In doing so, we explore a set of considerations relevant to the development of AI moral decision-making. Our main focus is on the relation between dual-process accounts of moral reasoning and model-free/model-based forms of machine learning. We also discuss how work in embodied and situated cognition could provide a valuable perspective on future research.

1 Introduction

Following recent theoretical developments in deep learning, researchers have started to consider how these technologies could be leveraged to help us understand the workings of the human mind (e.g. [51]). In a recent *Behavioural and Brain Sciences* article [30], however, Lake et al. argue that "despite rapid progress in AI technologies over the last few years, machine systems are still not close to achieving human-like learning and thought." Furthermore, they state that scaling-up current systems, or utilising more data, will not be sufficient to achieve human-like learning in AI, because fundamental ingredients of human cognition are currently missing. In this paper, we explore the proposal put forward by Lake et al.[3], focusing on a specific component of mind and intelligence they do not consider directly: our capacity for moral reasoning.

Our capacity for moral reasoning is influenced by the technologies we use. It will be further shaped by ongoing technological developments, such as those discussed by Lake et al. Therefore, building machines that "learn and think" like humans, as Lake et al. propose, raises important questions about the nature of morality and our capacity for moral reasoning. For example, how will our moral decision-making and deliberation be impacted when conducted alongside artificial moral agents? And, what directions should we pursue and avoid when designing artificial agents that learn and think (like humans) about morality? We discuss some of these questions, and look at possible research directions that we believe should be critically discussed by both philosophers and those directly engaged with the research and design of AI technologies.

Despite being speculative in nature, our article avoids considering extreme possible future scenarios (e.g. superintelligence), such as those made famous by the works of philosophers such as Nick Bostrom [3]. We believe many of the most important ethical challenges surrounding AI and morality are already upon us, and require careful interdisciplinary cooperation if we are to avoid the foreseeable quagmires inherent in our current and future moral landscapes. This is important to note, for it is far easier to construct a thought experiment concerning a distant future moral scenario than it is to plan for the way that AI will actually evolve. We believe there is good reason to focus on the present and the immediate future, and to take seriously proposals such as the one defended by Lake et al., as it is research such as this that gives us the clearest indication of what the future may have in store for us.[4]

The paper proceeds as follows. In section 2, we introduce the main claims defended by Lake et al., specifically their emphasis on the use of causal models in generalisable learning, and the distinction between model-free and model-based methods of learning. We briefly mention how these topics are to be connected to the discussion of moral reasoning. In section 3, we briefly discuss what aspect of moral reasoning we focus on, and give examples from the area of moral psychology to illustrate. We also connect this section's discussion to the distinction drawn in section 2, and explain why it is relevant to the prospect of building machines that learn and think about morality. In section 4, we introduce a more philosophical consideration about the representational requirements of internal model-building, and ask whether the proposal defended by Lake et al. could be further developed by considering work in situated and embodied cognition. We conclude, in section 5, by outlining a number of ethical issues that arise at the intersection of artificial intelligence and morality.

2 Building machines that learn and think like people

In their Behavioural and Brain Sciences target article [30], Lake et al. outline a research strategy, which they believe will help engineers to develop machines that "learn and think like humans". Their strategy focuses on three non-exhaustive[5], but core ingredients of human intelligence:

1. An ability to learn and build *causal models* of the world to support explanation and understanding, rather than merely solving pattern recognition problems.
2. Grounding this learning in *intuitive theories of physics and psychology* to support and enrich the knowledge that is acquired.
3. Harnessing compositionality and learning-to-learn to rapidly acquire and *generalize knowledge to new tasks and situations*.

[1] University of Bristol, Department of Computer Science, email: chris.burr@bristol.ac.uk
[2] University of Bristol, Department of Philosophy, email: gk16226@bristol.ac.uk

[3] Although we focus on the version of their proposal defended in [30], this account is an extension of the author's earlier work, known as 'Bayesian program learning' (see [29, 45])

[4] Nevertheless, we believe there is significant value in work such as Bostrom's. We simply choose not to adopt this strategy ourselves.

[5] In their author's response, Lake et al. acknowledge that many other faculties may also be required to enable machines to fully think and learn like humans, including emotions, embodiment and action, social learning and interaction, open-ended learning, and intrinsic motivation.

To demonstrate the importance of these ingredients they discuss two recent examples of state-of-the-art deep learning systems (see [29, 33]), which are trained on two separate tasks (i.e. handwritten character recognition and generation, and learning to play video games), but drastically differ from humans in terms of key performance indicators (e.g. poor transfer of domain-general knowledge; long training periods and large datasets). Lake et al. argue that current dominant approaches in machine learning are too entrenched in pattern recognition approaches, and fail to harness more human-like methods of learning, in order to transfer acquired knowledge to new domains. For example, they state:

> "A deep learning system trained on many video games may not, by itself, be enough to learn new games as quickly as people. Yet, if such a system aims to learn compositionally structured causal models of each game—built on a foundation of intuitive physics and psychology—it could transfer knowledge more efficiently and thereby learn new games much more quickly." (p. 18)

This idea reflects an assumption made by the authors that "the difference between pattern recognition and model building [...] is central to our view of human intelligence". As an example, they consider a video game called 'Frostbite'. This video game is notoriously hard for a typical deep learning (pattern recognition) system to learn [33], due to the need for long-term planning and complex hierarchically-structured goals (e.g. acquiring a series of items before a reward is offered). Even more recent versions of deep Q-networks, which eventually outperform a human player, require hundreds of in-game hours to achieve such performance. In contrast, most human players can achieve reasonable levels of performance in a matter of minutes. Furthermore, Lake et al. state that, once learned, a human player would be able to transfer prior knowledge about the game's causal structure to *novel scenarios* (e.g. novel game mechanics such as "Get the lowest possible score", or "Die as quickly as you can") very quickly. Importantly, these novel scenarios would represent drastic departures from the initial rewards learned from prior experience, and thus represent difficult hurdles for many deep learning systems designed through standard reinforcement learning techniques. Lake et al. argue that the knowledge transfer humans display likely relies on the existence of a constructed internal model, which represents a generalisable causal structure about the game's mechanics, and is leveraged by inductive biases inherent in human learning (so called "start-up software")[6].

In spite of the strong emphasis on model-based learning, Lake et al. also discuss model-free methods of learning. In reinforcement learning, a model of the environment is an optional element in an agent's *control policy*, where the policy is alone sufficient for determining behaviour [44]. Therefore, some artificial agents can act on the basis of model-free algorithms that directly learn a control policy without needing to build a model of their environment (i.e. reward and state transition distributions). However, such agents would require a model in order to undertake more complex forms of reasoning, such as long-term planning. As is well known in artificial intelligence, building a model of the environment can be costly and time-consuming, but as the above example highlights, model-free methods are inflexible outside of highly controlled domains, making them poor candidates for generalisable learning and knowledge transfer. Therefore, as Lake et al. argue, an agent that could make use of either cooperative or competitive mechanisms for switching between model-free and model-based forms of learning (see [12]), would appear to have an advantage over less flexible agents. Such an agent would also be closer to achieving more human-like forms of learning, as existing research suggests humans are capable of utilising both model-based and model-free methods of learning (e.g. [17, 38]).

This flexible switching between model-free and model-based forms of reasoning and learning is important for understanding how the above proposal connects to moral reasoning. In section 3, we will explore the application of dual-process theories of judgement and decision-making (e.g. [27]) to accounts of moral reasoning (e.g. [19, 47]). These theories claim that in addition to relying on deliberative, model-based forms of reasoning, human agents also rely on model-free heuristics that allow the agent to trade-off accuracy for speed, while potentially selecting value-enhancing actions in constrained environments [16]. Dual-process theories are ubiquitous in the sciences of human decision-making (e.g. behavioural economics), and are also common in evolutionary psychology where they are deployed as possible adaptationist explanations for a wide-range of observed behaviours in humans and primates [49, 14]. Prior to this discussion, however, it is important to address a theoretical assumption.

The perspective that Lake et al. adopt is explicitly computational in nature—that is, intelligent behaviour can be causally explained by appealing to a series of algorithmic processes that the agent's cognitive system realises [37].[7] However, Lake et al. also acknowledge, that this is unlikely to be sufficient to capture all forms of human intelligence:

> "Other human cognitive abilities remain difficult to understand computationally, including creativity, common sense, and general-purpose reasoning." (p.3)

In section 4, we will discuss a possible research avenue, inspired by work in social cognition and embodied robotics, which argues for the importance of cognitive scaffolds and niche-construction in supporting adaptive behaviour. We will argue that one possible hurdle that computational approaches could face, may arise with an implicit commitment to a methodological individualism, which views the brain's mechanisms as the primary system to be explained (in computational terms) when we wish to understand an agent's behaviour (see [9] for discussion). It is unclear to what extent Lake et al. are committed to a methodological individualism[8], and so our proposal is intended as a friendly suggestion that we believe is in the spirit of their account.

In contrast to the kind of methodological individualism that characterised classical cognitivist approaches to mind and behaviour, recent work in '4e cognition'[9] argues that some forms of human (and non-human) intelligence arise from an agent's engagement with its material environment (e.g. [32]) and embeddedness within its socio-

[6] Lake et al. acknowledge that human learning has itself been shaped by millions of years of evolution, which could be seen as our own "training period". However, this point merely reinforces their argument for developing a similar type of "start-up software" for artificial agents, which natural selection has developed for humans.

[7] Although we believe it is also worth considering whether moral reasoning could be better explained in non-computational terms, we restrict ourselves in this paper to considering research that is most directly relevant to the computational approach being discussed.

[8] For example, in their author's response, Lake et al. state that their intention was to remain agnostic about possible implementations for how models should be learned (see section R5.2 in [30]).

[9] '4e cognition' refers to work that fits within the research programmes of embodied, embedded, extended, and enactive cognition. It does not represent a unified research programme itself.

cultural niche (e.g. [2]).[10] The engagement between an agent and its environment can include the leveraging of physical structures (including the agent's own body) to reduce the computational demand that a given task places on the agent's cognitive system (e.g. reordering ingredients in a recipe, in order to reduce the demands on an agent's memory), but can also extend to normative constraints that are embedded within social institutions (e.g. language, legal structures, social norms). These normative constraints may themselves provide readily accessible alternatives to the costly construction of an internal model (e.g. using emotional feedback from peers as an approximate indication of whether your actions are socially acceptable).

In the following sections, we expand on each of the above points, which we believe offer fruitful ways of thinking about how to build artificial systems that could be capable of rudimentary forms of moral reasoning, or perhaps better support existing forms of human moral reasoning (e.g. "human-in-the-loop"). However, before we discuss these features, it is worthwhile stating what we mean by 'moral reasoning'.

3 What Is 'Moral Reasoning'?

When we ask what would be required for an AI to think and learn about morality, we must be clear about the kind of moral reasoning in question. There are, at least, three kinds of cognitive process which might reasonably be classed as 'moral reasoning'. In this section, we distinguish these different kind of moral reasoning, and make clear which of these is under consideration.

First, moral reasoning might be understood as the kind of reasoning demanded by the correct normative ethical theory (e.g. if utilitarianism is the correct theory, then moral reasoning is reasoning in accordance with utilitarianism). Second, moral reasoning might be characterised as a form of deliberation which requires us to adopt an *impartial perspective*. That is, moral reasoning requires us to consider the interests of a suitable reference class of moral patients, as opposed to just our own interests [41, 23]. Finally, according to a third, so-called *descriptive view*, moral reasoning might be understood as reasoning which involves *moral concepts*, such as fairness, duty, blame and responsibility. The focus of this third view is how humans do reason about morality, as opposed to how they ought to reason. As Lake at el. are primarily concerned with what is required to bring AI closer to human cognition, we focus on this third view.

The *descriptive view* involves a commitment to two claims. The first is that our moral concepts are built around innate tendencies to evaluate certain features of our environment [43, 26]. These evaluative tendencies admit evolutionary explanations. For example, take the disposition to evaluate characteristically *unfair* situations as bad. Plausibly, natural selection favoured genes promoting emotional dispositions *against* unfair situations, as these dispositions serve an important regulatory function that allows organisms to reap the benefits of prosocial behaviour. It is, therefore, no surprise that other primates have negative emotional responses to characteristically unfair situations [5, 6]. The second commitment involves the role of folk-psychological concepts in our moral reasoning. Guglielmo, Monroe and Malle [20], for example, have argued that many of our most important moral concepts are grounded in folk-psychology. Our concept of blame, for example, relies on our seeing the recipients of blame as *agents* capable of intentional action, foreseeing consequences, and so on. Joshua Knobe [28] defends a related thesis, according to which our moral concepts are central to how we understand intentional action. This commitment connects up with the first of the two insights from Lake et al., and we take it to be a positive feature of their argument that they acknowledge the relevance of grounding causal models in intuitive theories of folk-psychology in human cognition, insofar as this may help to capture important features of our moral reasoning. However, the types of models that Lake et al. emphasise are richly-structured generative models, which work by trying to reconstruct the hidden causal structure of a target domain (e.g. perception), and it is unclear to what extent this theory-like model-building is a necessary of our capacity for moral reasoning.

As alluded to in section 2, Lake et al. deal with this worry by reference to a key debate in reinforcement learning: the extent to which an intelligent agent relies on model-free or model-based methods of learning and decision-making. They acknowledge that some task domains are best approached using model-based methods of cognition (e.g. deliberative planning), whereas others seem to require model-free methods (e.g. skillful or habitual motor activity), and that some recent proposals in artificial intelligence and computational neuroscience use a combination of the two (e.g. [38, 48, 34]). The extent to which a particular task requires model-based or model-free methods of cognition is likely a matter of degree, and may require some sort of arbitration mechanism to alter the extent to which the two forms interact (see [10]). Regardless, neuroscientific evidence supports the idea that human learning comprises both model-free and model-based methods [17, 11]. How does this matter for moral reasoning?

If our moral concepts are grounded in our folk psychology, as Guglielmo, Monroe and Malle [20] argue, then one way of understanding the model-based versus model-free distinction, is as a guide to when our moral reasoning relies *most heavily* on deliberative or habitual processes[11]. Joshua Greene (e.g. [19, 18]) argues that moral reasoning involves an interplay between affective or 'quick-fire' cognitive processes and our deliberative cognitive processes, and this aspect of our morality may be nicely captured by the second insight from Lake et al. (i.e. an interplay between model-free and model-based learning). In his [18], Greene found empirical evidence showing that the way in which a moral dilemma is presented to us influences our deliberation about that dilemma. For example, in trolley cases, we are inclined to kill the one to save five if doing so involves causing the harm 'remotely' (e.g. pulling a leaver). But in cases which involve 'up close and personal harm' we are liable to have a negative emotional response which biases our deliberations in favour of letting the five die so that we avoid inflicting harm on the one.

This interplay can help us overcome some of the worst effects of using heuristic (or model-free) based forms of reasoning. As is well known, heuristics are often adaptive only in narrow domains [16], and there is some reason to think that heuristics make us *worse* moral reasoners outside of these constraints. Greene [18] and Peter Singer [40] have argued that the role of heuristics in moral reasoning causes us to be sensitive to morally irrelevant features of decision problems. For example, we are moved to help individuals suffering nearby to

[10] In some cases, the account that is offered rejects a computational perspective, in favour of a more dynamical approach (e.g. [8]).

[11] As already alluded to, it is likely that the extent to which certain forms of reasoning and learning are best described as "model-free" or "model-based" is a matter of degree. Therefore, it is ill advised to assume that the traditional dichotomies between habit and reason, or heuristics and deliberation, map neatly onto the distinction between model-free and model-based.

us, but not on the other side of the world, yet, according to Singer [40], the location of an individual is irrelevant to whether we ought to help them. In light of this, the proposal of Lake et al. to develop artificial systems that are able to adaptively deploy both model-free and model-based forms of learning and reasoning appears sensible in light of this worry.

In this section, we distinguished three different accounts of *moral reasoning* and specified the account which we intend to focus on. The descriptive accounts of moral cognition found in moral psychology will be the object of our inquiry. In what follows, we explore how model-free and model-based forms of learning, alongside embodied and situated cognition, can elucidate what it would mean for an AI to think and learn about morality.

4 The world as its "own best model"

In his [4], Rodney Brooks, offered a criticism of what he termed the 'sense-model-plan-act' (SMPA) model of artificial intelligence. The idea that Brooks wished to challenge was that if an AI (or a robot) was a) required to gather information from its environment (sensing), in order to b) build a richly reconstructive representation (model), with which to c) formulate a plan of reaching some desired goal-state (plan), before d) carrying out the necessary movements (action), then outside of a carefully designed and controlled laboratory setting (i.e. a narrow domain), such a serial process would be insufficiently dynamic to cope with the pressures of a constantly changing environment. In the time taken to deliberate, the environment may have changed, rendering the current model (and any actions based on it) inaccurate, and thus raising the agent's uncertainty. Utilising the SMPA model in ecologically-valid scenarios would mean either the artificial agent would incur an accuracy cost (subject to the environment changing), or it would incur a drastic speed cost. Instead, Brooks' suggestion was to implement a more straightforward sensorimotor coupling approach (based on his *subsumption architecture*), where the internal models were replaced with a more direct sensitivity to the environment, and the environment directly elicited and constrained adaptive behaviour with no need for mediating representations.

Since this time, greater consideration has been paid to the speed-accuracy trade-off, and the distinction between model-free/model-based methods has evolved to a point where many researchers now acknowledge the importance of some type of arbitration mechanism between the two methods (e.g. [10, 12]), rather than accepting a strict dichotomy. However, as some of the commentators to the Lake et al. target article argued, more attention still needs to be given to more ecologically-valid forms of intelligence that rely on the agent's situatedness or embodiment (e.g. [1, 35]). In short, if the body or environment of the agent enables the agent to offload some of the computational complexity, then there may be no need for the agent to construct a detailed inner model of the environment in the first place—in Brook's own words, "The world is its own best model." (1991, p. 15). This idea is reflected in work in developmental psychology [46] and soft robotics [36], and, in some cases, represents an instance of what Robert Wilson [50] refers to as 'wide computationalism'. It is also a familiar research area discussed in the 'extended mind' literature. In this section, we extend some of these considerations to the issue of moral reasoning—an area that is often underexplored in the 4e cognition literature.

One way of understanding the embodied and situated cognition research programmes, when applied to moral reasoning, is in uncovering the myriad ways that our environment (including our bodies) shapes and constrains the way we learn and reason about the world. Our environment represents an irreducible source of uncertainty and complex hierarchically-structured causes (e.g. what consequences will my actions have on other agents worthy of moral consideration), and our brains have clearly evolved heuristics and biases in order to simplify some of this complexity [16]. In acknowledging this, embodied and situated cognition researchers point to the way that social interaction allows us to cooperatively shape our sociocultural niche (e.g. [42]), and possibly make the world more predictable by constructing a more reliable domain in which our heuristics can operate (i.e. intervening on the world to reduce uncertainty). More recently, researchers in the area of *normative folk psychology*, have presented evidence for how certain sociocultural norms (including morality) are constructed through social interactions, and in turn contribute to our understanding of our own behaviour [52]. The benefit of constructing a stable, normatively structured environment is not only that it helps to regulate behaviour, but also that it provides a way of offloading some of the computational demands of cognition onto the environment itself. Acknowledging when this is possible (and desirable) could help AI researchers determine when artificial systems need to rely on model-based methods, or when the world can stand-in as "it's own best model". In the case of morality, by paying attention to the structure of the environment, engineers can determine if some normative structure is already present, and whether it is better to simply couple an agent's actions to the world as a sort of *distributed form of moral behaviour*. To better make sense of this, consider the following example.

H. L. A. Hart [24] provides an account of what distinguishes societies with a legal system from societies without a legal system. According to Hart, a set of norms becomes a legal system when 'secondary legislation' is enacted which stipulates the conditions under which a rule ought to be recognised as law. For example, the constitution of a state will specify which individuals are permitted to enact valid laws. If Hart's view is correct, then the development of secondary legislation could plausibly be construed as an instance of cognitive offloading with respect to moral cognition. Secondary legislation provides individuals in a society with a *prima facie* reason to behave in accordance with primary legislation, even if they do not understand the argument behind the primary legislation. Put another way, once secondary legislation is introduced, individuals have reason to comply with certain imperatives *because it is the law*. Thus, the existence of secondary legislation provides an external constraint on our legal (and often moral) behaviour, which does not require us to evaluate whether or not there is good reason to comply with the constraint.

The above example should highlight that a moral agent is not required to *internalise the norms* of society in order to ensure their behaviour meets certain moral standards, and can potentially make do with a simplified model of the world (or maybe even a set of well-tuned heuristics) when certain institutions act as regulative constraints. Of course, delineating the causal factors that govern an agent's behaviour is understandably a complex task. However, it is important to realise when an agent may be able to behave optimally (e.g. morally) simply by utilising adaptive heuristics, which respond to simple cues in the environment, rather than by constructing a rich inner model that acts as the basis for deliberative decision-making. This is not to deny that human agents are capable of norm internalisation, but the extent to which our moral behaviour is a product of constrained heuristics, rather than model-based deliberation is unclear. For example, it is possible that we achieve a high degree of moral optimality by using model-based reasoning (perhaps imple-

mented by mechanisms in prefrontal cortex) to competitively constrain the more heuristic based forms of action selection that drive our moral behaviour. However, it is also likely that society has collectively shaped our shared sociocultural niche, in order to reduce the demands placed on individuals, while nevertheless promoting optimal decisions.

An important question remains, why should we design artificial agents that rely on (potentially maladaptive) heuristic decision-making, like humans, when there is the possibility of pursuing more rational methods. Is there an answer, short of avoiding computationally demanding model-building, that can be offered?

5 Further Remarks

We conclude, by briefly considering whether the development of AI systems that are capable of human-like moral reasoning is a desirable goal. Our aim is not to argue exhaustively in favour of either positions, but rather to provide a sketch of the related ethical challenges that arise at the intersection of artificial intelligence and morality. We begin with the negatives.

5.1 Why artificial morality may be undesirable

(1) *The Ideal Reasoner Concern*: It might be the case that AIs which reason morally as *we* do are more inclined to make suboptimal moral decisions. As aforementioned, heuristics sometimes make us sensitive to morally irrelevant features of decision problems. So, if our intention is to develop AIs which make the *best possible* moral decisions, then we might have stronger reasons to focus on building *ideal* moral reasoners, as opposed to AIs which replicate our non-ideal moral reasoning. But, in order to design an ideal moral agent, we need to have a clear picture of what an ideal moral agent looks like [25, 3]. There are at least two problems here. On the one hand, there are several plausible ethical theories on the market, and moral philosophers are yet to provide decisive reason to favour one of these theories. Moral philosophers have only recently started to consider the rational response to 'normative uncertainty', providing decision-theoretic accounts of how to adjudicate between moral theories when we are unsure which, if any, is correct [31]. So, it is unclear at present which moral principles we have best reason to implement when designing an ideal moral agent. On the other hand, it is often the case that apparently plausible moral principles give surprising results in novel situations. Indeed, a substantial amount of ethical theorising involves testing how different principles square with our intuitions in novel cases. Whilst a set of moral principles might seem 'ideal' in one setting, they can easily be non-ideal in another. So, in developing AIs as 'ideal' moral reasoners, it is plausible that the principles used might deliver unforeseen and counterintuitive results, which is something we have good reason to be cautious about [3].

(2) *The Bias Concern*: Jonathan Haidt [22] notes the importance of in-group/out-group biases in our moral decision making.[12] Plausibly, whilst this bias may have an important functional role in human moral reasoning, there is good reason not to include biases of this kind in our AIs. The problem is that there exists a gap between how humans *in fact* reason morally given the available cognitive mechanisms, and how humans *ought to* reason morally. This gives rise to a trade-off. On the one hand, designing AIs to reason about morality as humans do will make AIs susceptible to the kinds of moral mistakes that humans routinely make. Human moral reasoning is imperfect, at least insofar as our cognitive mechanisms have inbuilt biases which dispose us to factor in morally irrelevant information into our decision-making. On the other hand, designing AIs to reason morally in a way that is too far removed from ordinary human reasoning will most likely result in AIs with inflexible moral reasoning that is unsuitable for general use across a broad class of moral decision-problems.

There are some biases in human cognition which would provide no straightforward benefit to AI moral reasoning if analogous biases were implemented into AIs. Furthermore, the non-inclusion of these biases is unlikely to be problematic. Consider *ego depletion*. According to what is called the *strength model*, humans have a capacity for self-control which enables them to engage in goal-directed behaviour and to resist impulses. However, this capacity is limited: prolonged activity involving self-control diminishes our capacity to resist impulses [21]. Although ego depletion no doubt plays a role in human moral decision-making, which is a taxing exercise requiring considerable self-control, there are good reasons not to include an analogous bias when developing AIs to reason morally. There is no clear benefit to having AIs which are in some way *worse* at making moral decisions after a prolonged period of moral decision-making. And it is desirable that AI moral reasoning is *consistent* over time. (This is especially important with respect to AIs making decisions which affect human wellbeing, such as AIs used to aid decision-making in criminal justice.) In our view, biases like ego depletion ought not to be considered as *necessary* for constructing an AI which reasons about morality, even though human moral reasoning is no doubt afflicted by this cognitive limitation.

We have examined two concerns about the desirability of AIs which have the capacity for moral reasoning. In the next section, we discuss two desirable features of AIs capable of moral reasoning.

5.2 Why artificial morality may be desirable

(3) *The Transparency Concern*: With regards to positive reasons for pursuing artificial morality, we first consider the issue of transparency in current deep learning systems [7]. The transparency concern relates to artificial decision-making systems that process information using methods that are opaque to most people affected by the AI's decision. It is, therefore, difficult to provide an *explanation* of how the AI reached its decision. Explanations are an important component of how we engage with other moral agents in society. Indeed, some moral theorists, such as T.M. Scanlon [39], argue that what makes actions morally right or wrong is whether those actions are mandated by principles which are justifiable to the parties affected by those principles. On this view, reasons take centre stage, as we justify our moral behaviour by appealing to reasons. In principle, the issue of transparency could be resolved if we develop AIs whose moral reasoning is grounded in folk-psychological mechanisms similar to our own moral reasoning. As Jonathan Haidt [22] and others have noted, when we attempt to justify our moral behaviour, these

[12] Note that there are three kinds of bias that present ethical issues in the context of AI decision-making: (1) We sometimes claim that a dataset is biased. When bias is a property of datasets, we mean that the sample data does not accurately represent the population. This kind of bias can present ethical issues with respect to, at least, training data for Artificial Neural Networks (ANNs). For example, an ANN used for voice recognition might be trained on a dataset of voices in which minority accents are insufficiently represented. (2) Other times we claim that a decision-making process is biased. For example, an ANN which is trained on biased data might produce skewed classifications. (3) Yet more times, we say that a decision-maker has a bias which is part of the agents cognitive apparatus. For example, an agent could have in-group/out-group bias as a component of their cognitive apparatus.

justifications involve folk-psychological concepts such as intention, belief and reasonable foresight of consequences. In bringing AIs in line with our mechanisms for moral reasoning, plausibly, this will open the possibility of AIs who can *themselves* offer moral justifications for decisions which are intelligible to those affected by the AI's decision. Importantly, as Daniel Dennett [13] notes in the case of the *Intentional Stance*, these types of explanations need make no reference to the underlying mechanisms that ground an agent's behaviour (e.g. the pattern of neural activity that causes an agent's actions), which is important given the inherent opacity of (black-box) deep learning systems. Explanations that are couched in terms of intentional psychological states (e.g. beliefs, desires etc.) play a simplifying role, which can also have a regulative effect on our future behaviour, and are typically sufficient to justify moral behaviour. For example, we are presumably happy for someone to justify their behaviour by virtue of appeal to folk psychological states, rather than a more complex explanation that makes reference to neural states[13]. Building artificial agents whose learning is grounded in intuitive folk psychological theories, as Lake et al. propose, seems a sensible first step in working towards artificial intelligence more understandable to humans.

(4) *The Envelope Concern*: Secondly, and finally, building artificial systems that can (co)operate within our own system of moral values is important, as we ideally want to avoid developing intelligent systems that are misaligned with our own moral principles. Research in situated and embodied cognition may represent a valuable avenue to explore in this regard, allowing us to develop autonomous decision-making systems that cooperate with us and help us overcome some of the limitations of our own moral reasoning. Luciano Floridi's [15] notion of an 'envelope' is a helpful conceptual tool to understand this point. He states, "In robotics, an *envelope* is the three-dimensional space that defines the boundaries that a robot can reach. We have been enveloping the world for decades without fully realising it." Here, the problem is that by "enveloping our world" such that it is easier for artificial agents to operate within it, we end up restructuring our own environment in ways that may have problematic consequences for us. We do not want our environment (including ourselves) re-structured to fit the ontology and values of artificial agents that may have conflicting goals or morals. In short, we want to design artificial agents that can (co)operate within our own envelope, not change our environment to fit theirs. Of course, the process will likely be a matter of reciprocal development (e.g. encoding knowledge systems in a AI-friendly manner; re-designing roads to accommodate autonomous vehicles), and humans are able to adapt to new situations thanks to our ability to learn generalisable knowledge. Pursuing artificial agents that "learn and think" more like us, however, may help make the process more conducive to human flourishing.

ACKNOWLEDGEMENTS

Christopher Burr is supported by a European Research Council Project [ThinkBIG, Advanced Grant (AdG), PE6, ERC-2013-ADG], awarded to Nello Cristianini. Geoff Keeling's contribution to this work was funded by an Arts and Humanities Research Council PhD Studentship, awarded through the South, West and Wales Doctoral Training Partnership.

[13] Of course, there may be exceptions to this. For example, a legal case in which a defendant appeals to an underlying defect in the neurophysiology, as an excuse for their behaviour. In these instances, folk psychological explanations may only be one component of the defendant's full justification.

REFERENCES

[1] Gianluca Baldassarre, Vieri Giuliano Santucci, Emilio Cartoni, and Daniele Caligiore, 'The architecture challenge: Future artificial-intelligence systems will require sophisticated architectures, and knowledge of the brain might guide their construction', *Behavioral and Brain Sciences*, **40**, (2017).
[2] Louise Barrett, *Beyond the brain: How body and environment shape animal and human minds*, Princeton University Press, 2011.
[3] Nick Bostrom, *Superintelligence: Paths, dangers, strategies*, Oxford University Press, 2016.
[4] Rodney A Brooks, 'Intelligence without representation', *Artificial intelligence*, **47**(1-3), 139–159, (1991).
[5] Sarah F Brosnan and Frans BM De Waal, 'Monkeys reject unequal pay', *Nature*, **425**(6955), 297, (2003).
[6] Sarah F Brosnan and Frans BM de Waal, 'Evolution of responses to (un) fairness', *Science*, **346**(6207), 1251776, (2014).
[7] Jenna Burrell, 'How the machine thinks: Understanding opacity in machine learning algorithms', *Big Data & Society*, **3**(1), 2053951715622512, (2016).
[8] Anthony Chemero, *Radical Embodied Cognitive Science*, MIT press, 2011.
[9] Anthony Chemero and Michael Silberstein, 'After the philosophy of mind: Replacing scholasticism with science', *Philosophy of science*, **75**(1), 1–27, (2008).
[10] Andy Clark, 'The many faces of precision (replies to commentaries on whatever next? neural prediction, situated agents, and the future of cognitive science)', *Frontiers in psychology*, **4**, 270, (2013).
[11] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan, 'Model-based influences on humans' choices and striatal prediction errors', *Neuron*, **69**(6), 1204–1215, (2011).
[12] Nathaniel D Daw, Yael Niv, and Peter Dayan, 'Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control', *Nature neuroscience*, **8**(12), 1704, (2005).
[13] Daniel Clement Dennett, *The intentional stance*, MIT press, 1989.
[14] Jonathan St BT Evans, 'Dual-processing accounts of reasoning, judgment, and social cognition', *Annu. Rev. Psychol.*, **59**, 255–278, (2008).
[15] Luciano Floridi, 'Enveloping the world for ai', *The Philosophers' Magazine*, (54), 20–21, (2011).
[16] Gerd Gigerenzer and Reinhard Selten, *Bounded rationality: The adaptive toolbox*, MIT press, 2002.
[17] Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P O'Doherty, 'States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning', *Neuron*, **66**(4), 585–595, (2010).
[18] Joshua D Greene, 'The secret joke of kants soul', *Moral psychology*, **3**, 35–79, (2008).
[19] Joshua D Greene, R Brian Sommerville, Leigh E Nystrom, John M Darley, and Jonathan D Cohen, 'An fmri investigation of emotional engagement in moral judgment', *Science*, **293**(5537), 2105–2108, (2001).
[20] Steve Guglielmo, Andrew E Monroe, and Bertram F Malle, 'At the heart of morality lies folk psychology', *Inquiry*, **52**(5), 449–466, (2009).
[21] Martin S Hagger, Chantelle Wood, Chris Stiff, and Nikos LD Chatzisarantis, 'Ego depletion and the strength model of self-control: a meta-analysis.', *Psychological bulletin*, **136**(4), 495, (2010).
[22] Jonathan Haidt, 'The new synthesis in moral psychology', *science*, **316**(5827), 998–1002, (2007).
[23] John C Harsanyi, 'Morality and the theory of rational behavior', *Social research*, 623–656, (1977).
[24] Herbert L A Hart, 'The concept of law', (1961).
[25] Michael Hauskeller, *Better humans?: Understanding the enhancement project*, Routledge, 2014.
[26] Richard Joyce, *The evolution of morality*, MIT press, 2007.
[27] Daniel Kahneman, *Thinking, Fast and Slow*, Macmillan, 2011.
[28] Joshua Knobe, 'Intentional action in folk psychology: An experimental investigation', *Philosophical psychology*, **16**(2), 309–324, (2003).
[29] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum, 'Human-level concept learning through probabilistic program induction', *Science*, **350**(6266), 1332–1338, (2015).
[30] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman, 'Building machines that learn and think like people', *Behavioral and Brain Sciences*, **40**, (2017).
[31] William MacAskill, 'Normative uncertainty as a voting problem', *Mind*, **125**(500), 967–1004, (2016).

[32] Lambros Malafouris, *How things shape the mind*, MIT Press, 2013.

[33] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al., 'Human-level control through deep reinforcement learning', *Nature*, **518**(7540), 529, (2015).

[34] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine, 'Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning', *arXiv preprint arXiv:1708.02596*, (2017).

[35] Pierre-Yves Oudeyer, 'Autonomous development and learning in artificial intelligence and robotics: Scaling up deep learning to human-like learning', *Behavioral and Brain Sciences*, **40**, (2017).

[36] Rolf Pfeifer and Josh Bongard, *How the body shapes the way we think: a new view of intelligence*, MIT press, 2007.

[37] Gualtiero Piccinini, 'Computationalism in the philosophy of mind', *Philosophy Compass*, **4**(3), 515–532, (2009).

[38] Evan M Russek, Ida Momennejad, Matthew M Botvinick, Samuel J Gershman, and Nathaniel D Daw, 'Predictive representations can link model-based reinforcement learning to model-free mechanisms', *PLOS Computational Biology*, **13**(9), e1005768, (2017).

[39] Thomas Scanlon, *What we owe to each other*, Harvard University Press, 1998.

[40] Peter Singer, 'Ethics and intuitions', *The Journal of Ethics*, **9**(3-4), 331–352, (2005).

[41] Peter Singer, *Practical ethics*, Cambridge university press, 2011.

[42] Kim Sterelny, 'Thought in a hostile world: The evolution of human cognition', (2003).

[43] Sharon Street, 'A darwinian dilemma for realist theories of value', *Philosophical Studies*, **127**(1), 109–166, (2006).

[44] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, volume 1, MIT Press, 1998.

[45] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman, 'How to grow a mind: Statistics, structure, and abstraction', *science*, **331**(6022), 1279–1285, (2011).

[46] Esther Thelen and Linda Smith, *A dynamic systems approach to the development of perception and action*, Cambridge: MIT Press, 1994.

[47] Michael Tomasello, 'Why be nice? better not think about it', *Trends in cognitive sciences*, **16**(12), 580–581, (2012).

[48] Théophane Weber, Sébastien Racanière, David P Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adria Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al., 'Imagination-augmented agents for deep reinforcement learning', *arXiv preprint arXiv:1707.06203*, (2017).

[49] Andrew Whiten, 'Chimpanzee cognition and the question of mental re-representation', *Metarepresentation: a multidisciplinary perspective. Oxford University Press, Oxford*, 139–167, (2000).

[50] Robert A Wilson, 'Wide computationalism', *Mind*, **103**(411), 351–372, (1994).

[51] Daniel LK Yamins and James J DiCarlo, 'Using goal-driven deep learning models to understand sensory cortex', *Nature neuroscience*, **19**(3), 356, (2016).

[52] Tadeusz Wieslaw Zawidzki, *Mindshaping: A new framework for understanding human social cognition*, MIT Press, 2013.

Symposium on Philosophy after AI: Mind, Language and Action

In conjunction with the 2018 Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2018)

6th April 2018

UNIVERSITY OF LIVERPOOL
RBC
BOREALIS AI
RBC Institute for Research
Taylor & Francis Group
an informa business