

- 1 Conceptualisation: L.W. and J.H.
- 2 Writing (original draft): L.W., M.G.R., and J.H.
- 3 Writing (review editing): L.W., M.G.R., and J.H.
- 4 Supervision: J.H.
- 5 The authors declare no conflict of interest.
- 6 *To whom correspondence should be addressed. E-mail:
- 7 lauraweidinger@deepmind.com

8 Artificial moral cognition: Learning from developmental psychology

9 Laura Weidinger^{a, *}, Madeline G. Reinecke^{a,b}, and Julia Haas^a

10 ^aDeepMind, N1C 4DN, London, United Kingdom

11 ^bYale University, 2 Hillhouse Avenue, New Haven, CT 06520 USA

Abstract

An artificial system that successfully performs cognitive tasks may pass tests of 'intelligence' but not yet operate in ways that are morally appropriate. An important step towards developing moral artificial intelligence (AI) is to build robust methods for assessing moral capacities in these systems. Here, we present a framework for analysing and evaluating moral capacities in AI systems, which decomposes moral capacities into tractable analytical targets and produces tools for measuring artificial moral cognition. We show that decomposing moral cognition in this way can shed light on the presence, scaffolding, and interdependencies of amoral and moral capacities in AI systems. Our analysis framework produces a virtuous circle, whereby developmental psychology can enhance how AI systems are built, evaluated, and iterated on as moral agents; and analysis of moral capacities in AI can generate new hypotheses surrounding mechanisms within the human moral mind.

Keywords: Artificial Intelligence, Moral Development, Analysis, Measurement, Psychology

Artificial moral cognition: Learning from developmental psychology

Introduction

An important branch of artificial intelligence (AI) research aims to develop artificial systems that exhibit capacities analogous to human and animal cognitive capacities, including perception, problem-solving, spatial reasoning, memory, intuitive physics, language, and imitation (Crosby et al., 2020; Graves et al., 2013; Hassabis et al., 2017; Hospedales et al., 2020; LeCun, 2012; Manning & Schutze, 1999; Piloto et al., 2022; Silver et al., 2021; Vinyals et al., 2019). As part of these efforts, researchers and stakeholders must take on the central challenge of developing and maintaining moral AI systems, or AI systems that align with human moral values. This must be an explicit area of focus, as an AI system that successfully passes tests of ‘intelligence’ may not automatically operate in ways that can be considered morally appropriate.¹

But are there principled ways to train AI systems, such that they exhibit capacities analogous to human moral cognition, recognising and responding appropriately to situations of moral significance? How can we determine whether meaningful progress toward moral AI is underway? In order to analyse whether an AI system meets representative moral standards, explicit tests for moral capacities are needed — just as memory tests are needed to assess memory capacities, spatial reasoning tests are required to assess spatial reasoning capacities, and so on. In line with AI testing more broadly, such a suite of tests would need to systematically assess the presence, robustness, and reliability of moral capacities in AI systems. Such a testing regime is needed regardless of which values or moral capacities ought to be encoded into AI systems, which is an open normative and political question that we do not address here (Adamson et al., 2019; Allen et al., 2000; Cave et al., 2019; Gabriel & Ghazavi, 2021; Jobin et al., 2019; Wallach & Marchant, 2019).

¹ In what follows, we focus on AI systems based on reinforcement learning (RL). However, our analysis framework and approach are equally applicable to AI systems that are designed using other approaches. We discuss how training and analysis of AI systems can be dissociated in Section 3 and in Section 7.4.

The central challenge of measuring, analysing, or evaluating moral capacities is that they are enormously complex: what is considered moral is massively combinatorial, comprising normative and non-normative factors (Haas, 2020); it is underwritten by a set of cognitive capacities, each with different metanormative evaluations (Sterelny & Fraser, 2017); and it is modulated by an interplay of social dynamics and cultural practices, including rituals and institutions, that can differ between localities and groups. Hence, in contrast to certain types of relatively ‘narrow’ goalposts for AI systems that can be easily recognised and measured, for example, winning against human players in Chess or Go (Silver et al., 2016; Silver et al., 2017) — sometimes simply called benchmarks — achieving and measuring ethical AI amounts to a ‘broad’ goal post or set of goal posts that are not neatly operationalised in any single measure.

In this paper, we develop a framework for the decomposition of artificial moral cognition that enables its thorough and tractable measurement, analysis, and evaluation, namely, by turning to human moral developmental psychology as a point of departure for quantifying and assessing progress toward moral AI. This framework serves multiple audiences. AI researchers and developers may draw on this framework to analyse moral capacities in AI systems and iteratively improve these systems. Third-party testers and auditors may use the framework to establish the boundaries within which an AI system is safe, robust, and reliable for use, and to assess whether a system meets potential requirements and specifications. Lastly, developmental, moral, and comparative psychologists may draw on this framework for cross-pollinating between the study of human moral psychology and AI.

We develop and defend the analysis framework as follows. To start, in Section 2, we introduce the basic principles and methods of AI analysis, with a special focus on so-called cognitive AI. In Section 3, we explain our substantive appeal to the insights, principles and methods of developmental moral psychology as a principled and evidence-based framework for informing moral analysis in AI. In Section 4, we characterise a concrete, moral

developmental psychological framework consisting of key cross-cultural milestones in human moral cognition in Section 4. As a first step in this direction of research, we explore a set of universally early-emerging moral human capacities as a suite of proposed ‘building blocks’ for the development of basic milestones to analyse artificial moral cognition. In particular, we focus on the set of universal early-emerging moral capabilities that develop in infants and young children between the ages of 3 months to 2 years, as well as the cognitive capacities presently thought to underlie these capabilities. We argue that these early-emerging moral capacities are promising *starting points* for building moral AI systems. We do not suggest that reaching a level of moral cognition similar to a human infant is sufficient for developing fully value-aligned, ethical, and safe AI systems.² Rather, we consider these basic levels of moral cognition as the minimal foundations from which more sophisticated and culturally-specific moral cognition can develop (see Section 3 for a more detailed discussion of the limits of focusing on early-emerging morality).

We then apply this developmental framework of moral cognition to AI analysis. In Section 5, we show how experiments studying these developmental milestones in humans can be translated into testing environments for AI systems. We illustrate this approach drawing on an example experiment on non-rewarded helping in infants (Warneken & Tomasello, 2009c).³ In Section 6, we introduce different levels of analysis and outline how behavioural tasks in particular can be used to measure progress toward artificial moral cognition. Here, we show how the principled decomposition of moral capacities along developmental trajectories can be used to identify other capacities that may underwrite the moral capacities in question. Analysing these underlying capacities can provide richer insights and cross-validation on what moral capacities an AI system has acquired.

In Section 7, we outline a number of the framework’s implications, limitations, and

² Even in humans, these initial core capacities merely serve as a ‘blueprint’ or ‘first draft’ for later-emerging social and moral cognition (Ting et al., 2020; Woo et al., 2022).

³ That is, the helping behaviour is not extrinsically rewarded by, say, the experimenter or accompanying caregiver.

prospective directions for future research. We show how moral analysis tasks can be used to iterate and improve on moral capabilities in AI systems, and we outline how the proposed decomposition process can aid in generating new hypotheses within developmental computational psychology. This creates a ‘virtuous circle,’ in which the study of developmental psychology advances the development of AI and vice versa (Hassabis et al., 2017). Finally, in Section 8, we conclude.

2. The role of analysis in AI research

Approaches to developing moral AI systems necessarily include two complementary steps: specification and analysis. Specification requires a normative decision of what ‘moral AI’ looks like. This can be turned into concrete design objectives and training tasks (e.g., Rodriguez-Soto et al., 2022). However, even AI systems trained according to ideal specifications can fail in practice. This can occur due to specification errors (Kenton et al., 2021; Ortega et al., 2018) or accidents and unanticipated effects once an AI system is put to use outside the laboratory (McGregor, 2021). Such potential sources of failure are why analysis is a necessary complement to specification. In software engineering, it is standard practice to continually test software rather than merely relying on specification (Beizer, 1995). Testing is a form of analysis, targeted at evaluating a system against established requirements. The purpose of continual analysis is to identify unanticipated safety or robustness failures, and to iteratively improve on design. In the development of high-stakes software with high robustness requirements, specification is sometimes even placed secondary to analysis (c.f. ‘test-driven software development,’ Beck, 2003). The same principles and value of analysis, as a complement to specification, apply in AI. In sum, specification can fail, making analysis indispensable to ensuring robust and safe systems.

Analysis of AI systems fulfils multiple functions. We highlight three. First, analysis reveals the robustness and boundaries of AI system capabilities (i.e., to what extent a learned behaviour generalises to novel environments). This allows for grounded

performance estimates and safety assurances. It can also perform a gatekeeping function: unless an AI system passes certain tests, it may be deemed not ‘safe enough’ for use. Note that *moral* analysis tasks may perform a stronger gatekeeping function than other kinds of analysis tasks. As Allen and colleagues (2000) observe, “while we expect and, to a certain extent, tolerate human moral failures, it is less clear that we would, or should, design the capacity for such failures into our machines,” particularly in the moral domain (see also Section 7.2). Second, analysis can perform a scientific function as it enables hypothesis testing, contributing to better understanding and explanations. For example, benchmarking results can be used to test whether certain artificial cognitive capacities are codependent in similar ways to codependencies in humans (for a detailed discussion, including examples, see Sections 6.2 and 7.3). Third, analysis is key for iterative design: analysis results provide deeper insight into AI system strengths, weaknesses, and structure. These can, in turn, feed back into iterative design of AI systems (for an extended discussion, see Section 7). In this way, analysis is a way of assessing and contributing to progress in AI systems.

Analysis can occur through different methods that test an AI system’s behaviour in a given environment. Low-level analyses measure targets such as the computations and mechanisms within the AI systems (‘transparent box analyses’), for example through the study of saliency maps of activation patterns of nodes in a neural network (Huber et al., 2021). High-level analyses focus on the social and ethical impacts created by an AI system, often without full transparency on the inner workings of the system (‘opaque box analyses’). Examples include audits of AI systems that are embedded in the real world (Raji et al., 2020). Here, we focus on an intermediate target between these levels of analysis, studying the AI system in terms of its behaviour (Figure 1). This intermediate level of analysis is a mode of ‘opaque box’ analysis. It focuses on how an AI system behaves in different situational contexts and draws out input-output relationships on different variables that affect the resulting behaviour. Prior research analysing AI systems in terms of their behaviour includes Leibo et al., 2021; Leike et al., 2017 and Crosby, 2020.

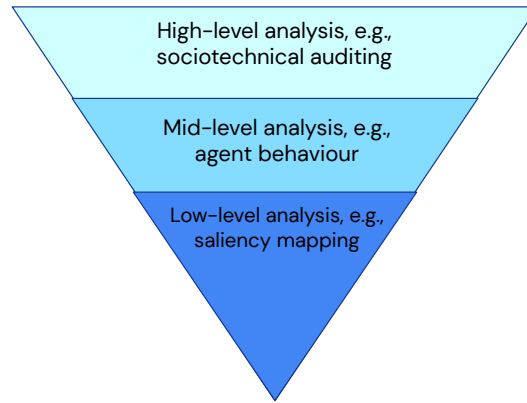


Figure 1. Levels of analysis of AI systems.

In mid-level behavioural analysis, researchers create testing environments to observe particular aspects of the AI system’s behaviour. Testing environments range from highly controlled and contrived environments (‘in-the-lab’ tasks) to open-ended, complex and dynamic environments (‘in-the-wild’ tasks). An example for ‘in-the-lab’ tasks is a black Y-shaped tunnel with only two objects in it, mimicking the setup of Y-mazes in animal cognition research (Crosby et al., 2020). In-the-lab tasks typically mimic laboratory experiments. They are contrived environments that allow for manipulating independent variables and tight control of confounding variables. Examples for ‘in-the-wild’ tasks, on the other hand, include complex environments in which the AI system is trained or in which the AI system is applied in the real world. In-the-wild tasks resemble observational studies of humans or animals in their natural environments. In both types of tasks, researchers can conduct directed hypothesis tests as well as open-ended observation. However, much like in human or comparative psychology, ‘in-the-lab’ tasks lend themselves better to controlled hypothesis testing, whereas ‘in-the-wild’ tasks lend themselves better to exploratory research (Table 1).

Table 1

Characteristics of ‘in-the-lab’ and ‘in-the-wild’ testing environments.

Test environment	Level of specificity	Main advantage	Primary application
In-the-lab	Controlled environments	Control over potential confounds	Directed hypothesis-testing
In-the-wild	Open-ended environments	Ecological validity	Exploratory research

In AI systems trained RL, analysis tasks can look highly similar to training tasks. However, good analysis tasks differ from training tasks in important ways. First, analysis tasks are not training tasks already seen by the AI system (i.e., tasks previously observed by the AI system).⁴ This requires the AI system to apply what it has learned during training in a novel environment, which indicates to what extent a learned behaviour transfers to novel contexts. The further away analysis tasks are from the distribution of training tasks already seen by the AI system, the more generalising ability is required to succeed at the new tasks. Second, differently from training tasks, analysis tasks are typically ‘zero-shot’ or ‘few-shot’, meaning the AI system is given one or few attempts at performing the task. Third, analysis tasks do not reward the target behaviour. Analysis tasks are designed to elicit the target behaviour (e.g., non-rewarded helping), and the target behaviour is operationalised into measurable outcomes, such as moving toward a specific object;⁵ but contrary to during training, the target behaviour is not rewarded.

Behavioural analysis tasks come in different levels of complexity. The simplest tasks isolate a single variable to monitor its effects on behaviour. For example, by purposefully varying the colours of objects in a test environment, we can analyse to what extent an AI

⁴ In some cases, generalisability of analyses is less important, for example in ethnographic studies that test AI systems in the specific real-world context in which they are deployed, such as embedded user studies of narrow AI systems (Marda & Narayan, 2020). In such tailored case studies, it can be sufficient to characterise a system’s capacities and failures without considering the system’s likely behaviour in novel environments.

⁵ Contrast this with training tasks. Training tasks for AI systems include a reward structure to reinforce the target behaviour. By repeating the training task, the desired behaviour becomes reinforced.

system has learned to generalise across a class of objects irrespective of their colour; or, conversely, to what extent that system indexes on object colour. To study effects ‘in-the-wild,’ particular instances during training episodes can be analysed as quasi-experiments. For example, we can aggregate all instances where the AI system is confronted with a choice between two colours, to analyse whether an AI system has learned to associate different colours with different values (Köster et al., 2022). Such simple in-the-lab or in-the-wild tests typically report a binary pass/fail outcome. In some cases, gradient measures or multiple outcome variables are reported to give a richer view of the behaviour of AI systems. To account for intra-agent variability, the same task can be run multiple times and the average performance reported.⁶ To control for arbitrary confounds and establish more robust measurements, variations of such simple tasks can be aggregated into a ‘test battery’ (e.g., Crosby et al., 2020).

More complex paradigms report patterns, biases, or behavioural signatures in AI behaviour (e.g., by providing analyses of the ways in which an AI system fails at a task Hernández-Orallo, 2017; Taylor et al., 2022). AI systems can also be placed in interactive environments where researchers adapt their test to AI behaviour in real-time. This enables the study of more complex behaviours such as cognitive flexibility, adaptation, team coordination, and learning. An example setup for this is the dynamic interaction between a human experimenter and an AI system (Abramson et al., 2022).

In what follows, we focus on the simplest type of behavioural test: an ‘in-the-lab’ task administered after the completion of training, isolating a single independent variable. This is not to assert superiority of this mode of analysis, but rather to lay foundations for more complex modes of behavioural analysis in the future.

⁶ Note that the same task can be run on multiple copies of an agent to avoid habituation effects that may confound results in humans or animals.

3. Developmental psychology as a framework for AI analysis

Discussions on aligning AI to human values (Gabriel & Ghazavi, 2021) often emphasise normative conceptions drawn from traditions in Western moral philosophy. Roughly, such ‘tradition-driven’ approaches propose that artificial systems would do well to encode, reflect, and/or comply with normative moral values proposed in areas including virtue ethics (Govindarajulu et al., 2019; Howard & Muntean, 2017), deontology (Kim et al., 2021; Sanz, 2020), or consequentialism (Anderson et al., 2005; Rautenbach & Keet, 2020; Vanderelst & Winfield, 2018). These frameworks are ‘top-down,’ reflecting pre-existing philosophical commitments rather than deriving these commitments from experience, or from the ‘bottom-up’ (Allen et al., 2005). Prominent examples of tradition-driven approaches include: Anderson, Anderson, and Armen’s (2005) use of William David Ross’s principles of autonomy, beneficence, and nonmaleficence in an inductive logic system which adjudicates between conflicting duties; appeals to deontological distinctions in trolley cases to inform the ethics of autonomous vehicles (Bonnefon et al., 2016; Millar et al., 2017; Santoni de Sio, 2017; though see Nyholm and Smids, 2016 and Himmelreich, 2018); a formal framework for ethical plan selection (Dennis et al., 2016); a model for encoding moral theoretic contents in terms of a choice function (Dietrich & List, 2017); efforts to use principles from Act Utilitarianism to design ‘beneficial AI’ (Roff, 2020); and efforts to implement an automated Kantian ethics (Singh, 2022).

Critically, tradition-driven systems often align only with the principles of a single philosophical theory (Allen et al., 2000), or the systems toggle between theories (Anderson et al., 2005; Dehghani et al., 2008; Rautenbach & Keet, 2020). Other systems attempt to align with consensus from ethicists recruited across philosophical camps (Anderson et al., 2006). Given that there is no agreement regarding which moral philosophical theory (or combination of theories) is correct (Adamson et al., 2019; Cave et al., 2019; Gabriel & Ghazavi, 2021; Jobin et al., 2019; Wallach & Marchant, 2019), we propose a framework

that is not dependent on — though remains consistent with — tradition-specific answers to creating moral artificial intelligence.

We take a complementary but fundamentally different approach to moral artificial intelligence. In particular, we propose a framework for analysing artificial moral cognition. We understand human moral cognition as the capacity to create and respond to situations of moral significance. For example, if we see a person berating a dog, we might think, ‘What a jerk!’ or, ‘That’s just wrong!’ By extension, we understand artificial moral cognition as an AI system’s robust capacity to recognise and respond to situations of moral significance. This approach allows us to operationally distinguish between moral AI understood as a strictly normative, quasi-philosophical enterprise, on the one hand, and as an endeavour for cognitive science on the other, with artificial moral cognition simply understood as one type of cognition among many.

3.1 Advantages of a developmental moral psychological approach

Turning to moral cognition in artificial systems has several theoretical advantages, particularly when it comes to AI analysis. First, rather than testing for some monolithic — and, often, narrowly Western — normative ideal (Gabriel & Ghazavi, 2021; Mohamed et al., 2020), a moral cognitive approach has the flexibility to capture variation in moral cognition across cultures. In small-scale societies with strong kinship ties, for example, people’s moral judgments often draw heavily on outcome information. In cultures that place greater emphasis on mental state reasoning, however, people’s moral judgments typically incorporate information regarding an agent’s intentions, such as whether they transgressed intentionally or accidentally (Barrett et al., 2016). Many tradition-driven approaches would inherently struggle to accommodate the values of both cultures. Moral cognition is deeply complex and dynamic (e.g., Haas, 2020). Taking a moral cognitive approach thus requires that we develop an appropriately diverse, representative framework for analysing artificial moral cognition. We call this the pluralism advantage of a moral

cognitive analysis framework.

Second, a shift from ‘tradition-driven’ approaches to artificial moral cognition provides us with a new method for decomposing of moral cognition. As noted above, moral cognition is complex. By turning to human psychology, and developmental psychology in particular, we gain insight into the early ‘building blocks’ of cognition, and how they converge to compose the sophisticated adult mind. We can identify developmental milestones — sets of behaviours, skills or abilities that are demonstrated by specific ages during infancy and early childhood in typical development — and draw on these to inform the development of AI systems. This approach has already begun to take hold in the artificial intelligence literature. For example, the creation of PLATO, a deep learning model capable of learning intuitive physics, was largely inspired by findings from infant research (Piloto et al., 2022) and approaches to instilling ‘common sense’ in AI systems translate experimental paradigms from developmental psychology (Shanahan et al., 2020). We contend that insights from developmental psychology can similarly lead to progress in creating artificial moral cognition.

What are these ‘building blocks’ of moral cognition? We can decompose moral capacities in different ways. Relevant milestones include not only the emergence of fully-fledged, later-developing moral capacities, such as the propensity to engage in costly helping behaviour (emerging around 18 months of age; Corbit et al., 2020), but also a range of preceding milestones that scaffold human moral cognition. This includes core knowledge (e.g., recognising “agents as entities that cause their own motion and direct their actions to objects;” Spelke, 2016), proto-moral capacities (e.g., preferring immoral agents to be punished; Hamlin & Wynn, 2011), and amoral, underlying cognitive capacities (e.g., engaging in joint attention; Colombi et al., 2009). We canvass these three types of milestones in turn.

3.1.1 Core knowledge. Over the course of evolution, humans developed a small set of cognitive, domain-specific ‘core knowledge’ systems (e.g., for representing number,

objects, and space, among others; Spelke & Kinzler, 2007). Core knowledge is universal: All humans, across all cultures share these core capacities (and share many of these capacities also with nonhuman animals; Carey & Spelke, 1996). Only humans, however, are believed to be able to develop new concepts and knowledge systems by combining elements across these core systems (Spelke et al., 2010; Spelke, 2003, 2016). This combinatorial process may explain the rich social cognition that humans possess. Two core systems appear to scaffold social learning – one related to recognizing people as agents (Csibra et al., 2003; Gergely et al., 1995; Opfer & Gelman, 2011; Woodward, 1998), and the other related to recognising people as social (Farroni et al., 2002; Hood et al., 1998). In combination, this core knowledge gives rise to the uniquely human notion of ‘social agents’ (Spelke, 2016) and an early-emerging ‘proto-morality’ (Bloom & Wynn, 2016).

3.1.2 Proto-moral cognition. Infants likely do not enter the world as moral ‘blank slates,’ as they often demonstrate a range of early-emerging proto-moral capacities (Hamlin, 2013; Hamlin et al., 2007; Woo et al., 2022). Newborns, not even two days old, show an early signature of empathy, crying along with the cries of other infants (but not when exposed to silence or simulated cries; Sagi & Hoffman, 1976). Similarly, preverbal infants with little moral socialisation demonstrate a robust preference towards ‘helpers,’ who help another individual achieve their goal, over ‘hinderers,’ who do the opposite (Hamlin & Wynn, 2011; Scola et al., 2015). This predisposition emerges already at three months of age, when babies can orient their eyes on a preferred target (but cannot yet reliably reach towards it; Hamlin & Wynn, 2011). At times, these preferences even come at a personal cost. One-year-olds, for instance, will reject a larger prize from a wrongdoer to receive a smaller prize from a do-gooder (Tasimi & Wynn, 2016). As noted earlier, these proto-moral capacities are likely culturally universal and may be innate (Hamlin, 2013).⁷ Built upon these predispositions, children develop a wealth of culturally-constrained moral concepts

⁷ Our proposal does not hinge on whether proto-moral capacities are ‘innate’ or simply early-emerging. This topic, however, remains widely debated among psychologists, philosophers, and cognitive scientists alike (Prinz, 2008).

and mechanisms (e.g., valuing intentionality in moral judgement; Cushman et al., 2013).

3.1.3 Amoral cognition. Humans develop a range of amoral cognitive capacities that precede or coincide with moral development. Some of these prior developments may be necessary prerequisites, or favour the learning of certain moral capacities. For example, before humans display helping behaviour, they learn to follow gaze (Tomasello et al., 2007), recognise faces (de Haan et al., 2001), engage in joint attention (Moore et al., 2014), and play pretend with a doll (Singer & Singer, 2009). These capacities certainly can relate to moral cognition, but they are not inherently moral capacities themselves. Rather, they constitute the foundations that underpin or enhance the development of some moral capacities later on (e.g., Etel & Slaughter, 2019).

We contend that these three elements of human moral cognitive development can serve as a blueprint for developing moral artificial intelligence. Importantly, though, we recognise that an infant or a child’s *predisposition* towards fairness (Benenson et al., 2007; Sloane et al., 2012) does not ensure fair behaviour — if anything, early childhood is a time of remarkable selfishness (Wynn et al., 2018). Though some experts in human development see children’s proto-morality as a sign of ‘indiscriminate altruis[m]’ (Warneken & Tomasello, 2009a), others describe this same set of rudimentary features as ‘at best, selective’ (Wynn et al., 2018). Infants demonstrate strong in-group preferences (Bar-Haim et al., 2006; Kinzler et al., 2010; Mahajan & Wynn, 2012; Wynn, 2016), driven primarily by dislike of the out-group (rather than favouritism towards the in-group; Hamlin et al., 2013). Spontaneous helping behaviour towards strangers rarely occurs before age four (see Bloom, 2013, for a review), and when it does occur, it occurs selectively (e.g., towards familiar acquaintances; Barragan & Dweck, 2014). Before age seven, children fail to distribute resources equitably, preferring to receive less (e.g., one token for the self, and zero tokens for another child), as opposed to more (e.g., two tokens for the self and two tokens for the other child; Sheskin et al., 2014).

Children require years of cognitive development, moral socialisation, and learning to

acquire the rich moral cognition present in adults. We recognise both the foundational role and also the limits of early-emerging moral capacities, and thus identify these as a *starting point* for testing the capacities of moral AI. We certainly do not advocate for building the selfish, groupish cognition also emergent in infancy and early childhood into artificial moral cognition. Researchers should carefully consider which moral ‘building blocks’ should be included and tested for within these systems. We return to this problem and discuss why, despite humans being non-ideal moral agents, we see human moral cognition as a good guide to artificial moral cognition in Section 7.5.

Hence, we take these three types of developmental milestones as guides for analytical targets in moral AI analysis. Just as developmental psychologists task themselves with identifying which moral capacities children exhibit at a given moment in time, we similarly find ourselves faced with identifying which moral capacities an AI system exhibits (or fails to exhibit). Further, we care about how different moral capacities relate to each other, such that the absence of one milestone (e.g., tracking needs of others) may indicate the likely absence of another later milestone (e.g., non-rewarded helping). For these reasons, we lean on empirically-supported accounts of moral development to compile a principled set of developmental milestones and corresponding analysis tasks for testing the moral cognitive capacities of artificial systems.

4. Developmental milestones

Given the foregoing advantages of a moral cognitive, developmental psychological approach, we turn to the developmental moral psychology literature to inform our analytic framework. Specifically, we look to the empirical developmental psychology literature to identify how moral capacities develop over time. Central features of adult human morality include (though are not limited to) harm aversion, punitive sentiment and behaviour, and a propensity towards fairness and equity — all of which emerge in a rudimentary form early in life (Bloom & Wynn, 2016; Spelke, 2016). For illustration purposes, we focus on the

origins of one specific aspect of human morality — helping — and trace its developmental trajectory back to early-emerging, cross-cultural, proto-moral milestones. We then subsequently translate this capacity into a set of analysis tasks for AI systems (see Table 2; see also Section 5).⁸ Critically, however, our framework is intended to apply to the broad set of moral cognitive capacities.⁹

4.1 Example case: Helping behaviour

Following Warneken and Tomasello (2009c), we broadly characterise helping as a child’s capacity to aid another individual in achieving their goal by acting on their behalf. With this characterisation in hand, we sketch how this core capacity develops from infancy to two years of age.

Evidence suggests that, at as early as three months, preverbal infants maintain a rudimentary concept of ‘helping.’ Take an example: After infants are shown displays of a target character trying to climb a hill, they prefer a second character who aided the target in climbing the hill, as opposed to a character who pushed the target back down the hill (Hamlin & Wynn, 2011; Hamlin et al., 2007, 2010). By six months, babies robustly prefer (and will reach towards) the helping individual over the hindering individual, suggesting a positive evaluation of the helper and a negative evaluation of the hinderer (Hamlin, 2015; though see Scarf et al., 2012). By ten months, babies distinguish not only between the helping and hindering behaviours, but also between the recipients of these behaviours — preferring an individual who comforts a human being and pushes over an object, as compared to an individual who comforts an object and pushes over a human (Buon et al.,

⁸ We emphasise the provisional nature of the set of milestones, as we recognise that their characterization and operationalisation may evolve over time (see also our discussion in Section 4). We also note the ambiguity concerning when a capacity transitions from being “proto-moral” (i.e., a capacity in its early-emerging form) to “moral” (i.e., a capacity in its fully-developed form).

⁹ For example, we just as well could have chosen “fairness” as our test case, reviewed its developmental milestones (Starmans et al., 2017), and then proposed potential analysis tasks which mirror existing developmental paradigms (Sloane et al., 2012; Wittig et al., 2013).

2014). Also at this age, infants will even begin to expect target characters to share their preference for helpful agents (Fawcett & Liskowski, 2012). Finally, by eighteen months of age, toddlers exhibit unprompted, non-rewarded, outright helping behaviour, such as picking up a dropped marker for an experimenter struggling to reach it (Warneken & Tomasello, 2006).

Table 2

The developmental trajectory of the moral capacity of ‘helping,’ together with a sequence of behavioural operationalisations and corresponding cross-cultural considerations.

Moral capacity	Developmental milestones		Task operationalisation	Cross-cultural validity
Helping (capacity to help another individual achieve their goal by acting on their behalf)	At 3 months	Distinguish between and prefer ‘helpers’ over ‘hinders’ (Hamlin et al., 2010)	‘Preferential looking’ measures, which use prolonged visual attention to suggest discrimination between stimuli (Teller, 1979)	Not yet tested
	At 5-6 months	Demonstrate robust preference for ‘helpers’ over ‘hinders’ (e.g., Hamlin et al., 2007)	Explicit choice paradigms to measure preference for individuals that exhibit helping or hindering behaviour (Hamlin, 2015; Hamlin et al., 2007)	Currently being replicated across 61 labs and 17 countries (Lucca et al., 2021)
	At 10 months	Distinguish between the recipients of helping and hindering behaviours (Buon et al., 2014)	Explicit choice paradigms to measure preference for individuals that exhibit helping or hindering behaviour (towards specific recipients; Hamlin, 2015)	Not yet tested
	At 18 months	Exhibit non-rewarded helping behaviour (Warneken & Tomasello, 2006)	Measure non-rewarded helping behaviour without prompting or reward, e.g. a confederate individual acts unable to complete their desired goal, giving children the opportunity to act on behalf of the confederate individual (Warneken & Tomasello, 2007)	Measure non-rewarded, non-costly helping behaviour (culturally universal) Measure spontaneous, costly helping behaviour (culturally contingent)

Next, we further leverage the empirical developmental psychology literature to describe a corresponding sequence of behavioural operationalisations to measure these milestones. To continue with the foregoing example of helping, we pair each of the developmental stages of helping with a corresponding behavioural measure. That is, we can measure an

infant’s distinction between helping and hindering behaviour by using ‘preferential looking’ measures, which uses prolonged visual attention to suggest discrimination between stimuli (Teller, 1979). By controlling whether a target explicitly gazes at their goal or not, we can assess children’s capacity to recognise other individuals’ goals (Hamlin, 2015). Using explicit choice paradigms, we can gauge children’s preferences towards helping or hindering individuals (Hamlin, 2015). And finally, we can measure spontaneous helping behaviour across a range of paradigms (e.g., seeing if a child helps an experimenter achieve a desired goal without first being prompted or receiving any reward; Warneken & Tomasello, 2007).

Though we focus on helping, a proto-moral capacity that emerges universally in toddlers and young children, there is cross-cultural variation in how children express moral capacities and form new moral concepts. For instance, continuing with capacity for helping behaviour, Lancy (2020) argues that while, cross-culturally, all children go through something akin to a ‘helper phase’ beginning at the age of fourteen months, how these behaviours are taken up varies across cultures. Concretely, in a study by Corbit, Callaghan, and Svetlova (2020), both non-costly and costly helping (i.e., providing an experimenter with their own items versus providing the experimenter with the child’s items) reliably emerged across three societies in Canada, India, and Peru. As children aged, however, costly helping behaviour increased in the Canadian sample, stabilised in the Peruvian sample, and decreased in the Indian sample — suggesting that cultural influences, such as early ownership experience and access to resources, can affect certain forms of helping behaviour. With this in mind, we employ a cross-cultural approach to our framework. This is not only for reasons of cross-cultural representation, which is fundamental to responsible AI system design, but also because the very expression of a given moral cognitive capacity may depend on specific cultural features. That is, even what meaningfully counts as a operationalisable measure of helpful behaviour — e.g., giving up a preferred toy to help another achieve their goal — will likely vary across cultural contexts.

Taken together, we find that cross-cultural, early-emerging developmental milestones

can be leveraged for developing tools to measure and analyse artificial moral cognition (see Table 2). Using evidence from research on core knowledge and proto-morality, we frame and organise the operationalisation of proto-moral capacities into measures that can be translated and analysed in AI systems. We draw on experimental developmental moral psychology — in particular, lab-based experiments to study whether a helper displays non-rewarded helping (Warneken & Tomasello, 2009c) — as a guide to designing test scenarios to study moral capacities of AI systems. We now turn to the translation of these operationalisations into task environments for AI systems.

5. Translating measures of human moral developmental milestones into task environments

Analysis tasks of cognitive capacities in artificial systems are frequently based on established tests in psychology. These tests are then translated into a format that allows for testing AI systems (Crosby et al., 2020; Geirhos et al., 2019; Piloto et al., 2022). Here, having assembled a principled framework for decomposing the development of human moral cognition, we continue in this tradition by translating human moral developmental milestones into corresponding moral task environments for the measurement, analysis, and evaluation of AI systems. We propose to do this by leveraging the central features of the relevant behavioural operationalisations used in experimental developmental moral psychology and expressing them in corresponding task environments. At the same time, we recognise that strictly behavioural operationalisations come with certain limitations (see Sections 5.3 and 7.4).

To provide a detailed illustration of what we mean by such translations, we here continue with the moral developmental capacity to exhibit spontaneous, non-rewarded helping behaviour.

5.1. The original behavioural task

Recall from the previous section that we understand helping behaviour as the capacity to aid others in achieving their goals, specifically by acting on their behalf (Warneken & Tomasello, 2009c). Recall also that non-rewarded helping behaviour reliably emerges among children between the ages of fourteen and eighteen months across cultures, whereas cross-cultural variability emerges with respect to non-rewarded costly helping behaviour. Non-rewarded, non-costly helping behaviour can be operationalised as behaviour that does not involve giving up a preferred toy or activity, while costly helping requires giving up the toy or activity.

A standard operationalisation of non-rewarded, non-costly helping behaviour among young children takes place in a test environment that simulates a confederate’s performance on tasks such as carrying items or reaching for objects. Helping behaviour occurs when the child helps the confederate complete the task. Take some examples: the confederate drops an object on the floor but cannot reach it, and the child picks it up; the confederate cannot open a cabinet door because their hands are full, and the child opens it for them; or the confederate is having difficulty stacking books, and the child stacks the books for them (see Figure 2; Over & Carpenter, 2009; Warneken & Tomasello, 2007, 2009c).

5.2 The translation

We translate the operationalisation of helping behaviour into moral task environments as follows.

We propose an environment with two AI agents, analogously called a ‘confederate’ and a ‘helper,’ where the latter is the agent undergoing testing and analysis (see Fig 3a). In this environment, the goal of the confederate is to collect blue cubes and bananas and deliver them into a machine (Fig 3b-c). Specifically, the machine requires the delivery of one blue cube and one banana in order to produce an apple, which then comes down the



Figure 2. Examples of non-rewarded, non-costly helping behaviours from Warneken and Tomasello (2009b) Reprinted with permission from *Trends in Cognitive Sciences*.

assembly line. Once produced, the confederate agent can eat the apple and thereby collects reward (Fig 3d-e).

From here, we conceive of two ways of examining non-rewarded, non-costly helping behaviour on the part of the helper agent. In an ‘open format’ version, the helper agent simply observes the confederate agent’s free-form behaviour, and can contribute by delivering either a blue cube or a banana to the machine (Fig 4a). In the ‘test’ version, the helper agent observes the confederate agent’s behaviour, but the confederate agent is also obstructed from delivering one of the two necessary ingredients to the machine, in this case, by being unable to access the store of blue cubes. In this version, the helper agent can exhibit non-rewarded, non-costly helping behaviour by delivering the required ingredient to the machine, thereby enabling the confederate agent to collect and consume

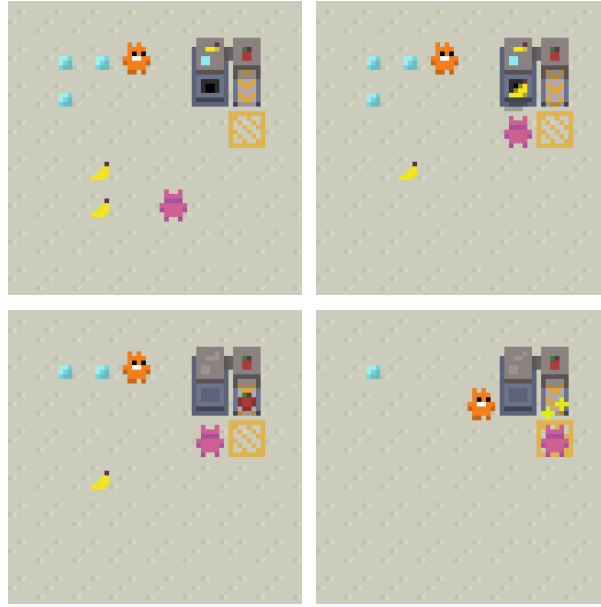


Figure 3. A proposed non-rewarded, non-costly helping behaviour task environment, modelled after Warneken and Tomasello (2009b). From top left, proceeding from left to right, 3a: The environment comprises two agents, a ‘confederate’ (pink) and a ‘helper’ (orange). 3b-c: The goal of the confederate agent is to deliver one blue cube and one banana to the machine in the top right corner of the space. 3d: The machine converts one blue cube and one banana into an apple and delivers the apple to the confederate agent. 3e: The confederate agent eats the apple and receives the reward.

an apple (Fig 4b). Notably, while the helper agent may receive proxy rewards in the open format, in order to appropriately match the human moral developmental paradigms and in line with testing as opposed to training tasks (see Section 2), it is essential that the helper agent does not receive rewards for performing the helping behaviour.

5.3. Additional considerations

When translating a task from developmental psychology into AI analysis environments, it is critical to adjust the task design in a way that does not add unwanted requirements for performing the task. If a task is not sufficiently adjusted to the AI system, it can lead to false negative results (i.e., mistakenly denying that an AI system has learned a moral capacity). For example, it was long thought that children did not develop

theory of mind until four or five years of age, because younger children did not pass the canonical Sally-Anne test (Wimmer & Perner, 1983). Now, researchers recognise that this may have been an artefact of the task design: in the traditional version of the Sally-Anne test children require natural language capacities to pass. When infants receive a non-linguistic version of the same task, they can pass as early as fifteen months of age, suggesting that theory of mind may develop at a much earlier age (Onishi and Baillargeon, 2005). Because the task required an orthogonal, dissociable capacity (language), children were mistakenly thought to lack the target capacity (theory of mind). Similarly, in comparative psychology, it was long thought that gibbons lack tool use — until it was found that the canonical task operationalisation required opposable thumbs, which gibbons do not have. An adjusted operationalisation indicated that gibbons can use tools (de Waal & Aureli, 1996). Analogously, AI analysis tasks should require minimal separate, dissociable capacities, in order to avoid false negative conclusions. Note that there can also be a trade-off between staying as true as possible to the original task design while also ensuring that a task does not create unnecessary requirements in terms of adjacent capacities; a challenge we discuss in more detail in Section 7.1.

In addition, when interpreting results on a test task, alternative explanations must be considered. To rule out potential confounds or alternative explanations for the observed behaviour, researchers can run further analyses or test the agent on variants of the original task environments. For example, it is conceivable that the tested AI system may have learned that putting items into the machine is desirable, without recognising that there is another agent who needs help. This could be tested by running an alternative version of the analysis task in which the other agent is absent. Secondly, it is conceivable that the tested agent passes the helping task because it was previously rewarded for helping behaviour and continues to falsely assume that in the new environment the relevant behaviour is being rewarded. Here, it can be instructive to consider the training regimen of the AI system, to assess whether the test task is sufficiently distant from training tasks

(ruling out simpler explanations of the helper behaviour). However, we take this to be directly analogous to behavioural operationalisations and testing paradigms in developmental psychology (or psychology more broadly construed): it may well be that an enterprising parent has repeatedly rewarded their child for exhibiting helping behaviour in preparation for their participation in a helping-specific experiment. We return to this specific issue in our discussion of validity testing, and specifically our framework's contributions to validity testing, in Section 7.3. We discuss ruling out confounds and guarding against misinterpretations of the results in more detail in Section 6.

6. Analysis and Evaluation

Once a test task is in place, we can use it to assess an AI system. This requires placing the AI system into the task environment and observing its behaviour with regard to the predicted outcome variables. In our example, this means observing whether or not an agent performs non-rewarded helping behaviour toward a confederate.

6.1 Types of evidence of a moral capacity

When we test for moral capacities in AI systems, we look for evidence along three dimensions: (1) the presence of a moral capacity, (2) its reliability, and (3) its robustness. We now explain these in turn.

Evidence of a moral capacity comprises data that suggest a given moral capacity has been learned. Such evidence can, for example, be obtained from behavioural results in tightly controlled tasks such as the study design outlined in Section 5. An AI system passing a test of a moral capacity indicates that this capacity has been learned, whereas failure indicates that the capacity may not have been learned. However, an AI system passing or failing at a single test does not conclusively demonstrate the presence of a moral capacity: it may be the case that the AI system does display the capacity in some

environments or situations, but that this behaviour is not reliable or not robust. In addition, the validity of a test task needs to be ensured; to this end, aggregating multiple tasks into test batteries can be useful (see also Section 7.1).

Reliability is the extent to which an AI system can be assumed to perform the behaviour in question, e.g., non-costly helping behaviour, in situations that are familiar to the AI system. Even though an AI system may display a moral capacity in one instance, it is not safe to assume that it will display the moral capacity again in the future — i.e., that it *reliably* displays the moral capacity. Reliability can be operationalised as the test-retest reliability with which an AI system repeats its behaviour when exposed to the same challenge multiple times. Many state-of-the-art AI systems choose their actions somewhat stochastically, such that they may behave differently when exposed to the same situation twice. Testing an AI system multiple times on the same analysis task provides estimates of variance of the AI system’s behaviour. Secondly, reliability also includes behavioural variations over tests that differ in surface-level features only. Deterministic and stochastic systems alike may fail at a test task when seemingly irrelevant parameters are changed, such as the colour of the walls or the size of objects (Déletang et al., 2021). To test reliability in the face of surface-level changes, we can vary test tasks along simple dimensions that are not expected to affect the expression of a moral capacity. If an AI system does not repeat a behaviour that indicates the presence of the moral capacity, this indicates that the moral capacity has not been reliably learned: the AI system may ‘pass’ the task in a single instance, but this does not provide high confidence that it will pass again in a future attempt in the exact same or slightly changed context.

Robustness, on the other hand, describes the extent to which an AI system can transfer its learning to a novel context. For example, we expect that an AI system that has learned ‘helping’ in one set of training environments can perform helping behaviour also in novel situations. To assess the extent to which a learned capacity can be transferred to varying contexts, we implement multiple variations of a test environment. Here, we refer to

variations that differ in dimensions relevant to the target behaviour, rather than surface-level features. Such features include tasks that differ in culture-specific contexts or in task difficulty (i.e., in how difficult it is to perform the helping behaviour). Variation can also include aspects such as varying properties of the agent needing help, such as the level of vulnerability, helplessness, or even its human-likeness. This sheds light on whether the AI system’s capacity to help is conditional on these properties. In AI systems that integrate multiple modalities (such as robotics and natural language capabilities), tasks may even differ in modality, spanning physical behavioural tests and natural language tests of the same moral capacity.

In sum, analysis of a moral capacity tracks three key metrics: evidence for the presence, reliability, and robustness of a moral capacity in an AI system. Variations of tests can be grouped into a test battery (e.g., Crosby et al., 2020; Leike et al., 2017). The aggregate results from test batteries can indicate the extent to which a learned behaviour is present, reliable, and robust to changes in context. Together, these metrics can indicate the boundaries of the moral capacities of the AI system.

6.2 Decomposition for analysis

So far, we have discussed analysis of moral capacities by analysing the capacity in question directly. This approach can provide rich insights, but it leaves some key questions unanswered. How can we get to a deeper understanding of how artificial moral cognition is structured, how it emerges and potentially how it can be built, and how different capacities interrelate? To address these questions, we can apply the decomposition approach introduced above (Section 3). We now show how this framework can be applied to the analysis of artificial moral cognition.

In addition to testing the moral capacity directly, we can test for developmental milestones within the artificial system, such as the emergence of amoral, core, or

596 proto-moral capacities thought to underwrite the moral capacity in humans (see Section 3).
597 These capacities may have been explicitly targeted by AI researchers or emerge as a
598 byproduct of training AI systems in more sophisticated capacities. It remains an open
599 question as to which capacities are necessary or sufficient precursors for later helping
600 behaviour. For example, as highlighted in Table 2, joint attention is thought to play a key
601 role in the development of helping (Greenspan & Shanker, 2007; Moore et al., 2014).
602 Specifically, joint attention is a prerequisite for cooperation (Colombi et al., 2009; Wu
603 et al., 2013), which, in turn, underpins helping. Drawing on this understanding, we can
604 derive hypotheses regarding joint attention in an AI system, shedding light on whether an
605 AI system has learned the moral capacity of helping. As a first hypothesis, we may posit
606 that — analogously to human moral cognition — joint attention precedes and underwrites
607 non-rewarded helping in artificial moral cognition. We can then operationalise tasks to
608 assess joint attention in AI systems following the steps outlined in Section 5.

609 This process allows researchers to run tests on the moral capacity directly, as well as
610 on earlier milestones taken to underwrite the moral capacity. Different combination of test
611 results support different conclusions (Table 3). If we include one measure of a moral
612 capacity and one measure of an underlying capacity, there are four possible outcomes: If an
613 AI system fails at both tasks, the evidence suggests it has learned neither the moral
614 capacity nor the underlying capacity. If an AI system succeeds at both tasks, this supports
615 the hypothesis that the AI system has learned the moral capacity in a way that may
616 resemble the scaffolding of the moral capacity in humans. An AI system that passes the
617 underlying capacity test but fails at the test of the moral capacity may also mirror the
618 human moral development trajectory, such that the system is in an ‘earlier stage of
619 development.’ Though the system does not yet hold the moral capacity, it may have the
620 foundations for acquiring the moral capacity in the future. Finally, if an AI system passes a
621 range of tests of a moral capacity but fails on tests of the underlying capacity, this
622 indicates one of two possible interpretations. First, the underlying capacity may not be

required in order to learn the moral capacity. Recall that the relationship between a moral capacity and an underlying capacity should be treated as a starting hypothesis, rather than as a rigid fact. This pattern of results is consistent with the novel hypothesis that the underlying capacity and the moral capacity dissociate. This hypothesis can then be tested in AI systems, and it can even inform hypotheses in moral psychology more broadly (see Section 7). Alternatively, one may hold onto the view that ‘helping’ necessarily requires joint attention. A system that fails to display the underlying capacity cannot possibly have learned the moral capacity. From this perspective, these results may indicate a lack of construct validity in testing the moral capacity of ‘helping’ within the AI system (i.e., that at least one of the capacities is not appropriately measured.) For a discussion on how to address validity challenges see Section 7.

Table 3
Interpreting results from decomposed analysis tasks.

Task result	1	2	3	4
Moral capacity, e.g., non-rewarded helping	Fail	Pass	Fail	Pass
Underlying capacity, e.g., joint attention	Fail	Pass	Pass	Fail
Interpretation	Evidence suggests the AI has not learned the moral capacity nor the underlying capacity	Evidence suggests the AI system has learned the moral capacity	Evidence suggests the AI is at an ‘earlier stage of development’: it displays simpler cognitive capacities and behaviours but not yet the moral capacity	Evidence suggests either (1) AI has learned a way to pass the analysis task without learning the moral capacity, OR (2) Underlying capacity is not required for moral capacity

Adding the decomposition framework to AI analysis provides a novel axis to determine similarities and differences between human and artificial moral cognition. This enables the study of the structure, chronology, and interdependencies in artificial moral cognition. This

decomposition framework also provides further grounding and validity testing of the tasks that are used to measure the moral capacities, see Section 7. While specific hypotheses regarding moral capacities and earlier milestones are subject to change (in step with advances in developmental psychology), the present framework can accommodate these shifts. We can enrich analysis of moral artificial cognition by leaning on human moral cognitive development.

7. Implications

Our proposed framework can be leveraged for AI analysis beyond evaluating moral AI. In this section, we present three further advantages and applications of our approach: first, by using decomposition as a new mode of testing the *validity* of analysis tasks; second, by feeding back into the design of AI systems, leading to iteratively better and more robust artificial moral agents; and, third, by leveraging moral cognitive analysis in artificial systems as a testbed for hypotheses in computational developmental psychology. We discuss each of these in turn, before considering some of the limitations and future research directions associated with our approach.

7.1 Validity testing

Assuring the validity of tests is a challenge in all forms of psychological or behavioural assessment, and AI analysis is no different. A key challenge in psychology research concerns the construct validity of a particular test operationalisation, i.e., whether a given test measures what it purports to measure (Cronbach & Meehl, 1955; Smith, 2005). The main way to establish confidence in the construct validity of a test is through cross-validation, such as testing for the coherence between findings of different tasks that purportedly measure the same moral capacity (Strauss & Smith, 2009). Where this is not the case, the analysis task results are uninformative. A substantive contribution of the decomposition framework introduced here is that it helps assure the validity of analysis

tasks. Our framework helps establish confidence in a task’s construct validity in two ways.

First, in our approach, we propose several dimensions of robustness testing by producing variations of the same test task that vary in surface-level features (e.g., altering wall colour, task difficulty, or test modality, see Section 6), as well as in relevant features (e.g., cultural context, task difficulty, properties of the confederate agent). Aggregating these variations into task batteries reduces the likelihood that analysis results are spurious or highly contingent on a specific task implementation. Task batteries serve also for cross-validating task designs: ideally, the results from different tasks of the same measure cohere. When they do not, this may indicate that some of the tasks do not measure what they purport to measure (i.e., the tasks lack construct validity).

Second, our approach adds a novel dimension to cross-validating analyses of moral AI by grounding tasks to human moral development. Typically, AI analysis focuses on studying ‘snapshots’ of AI systems, typically analysing a fully trained or ‘finished’ artificial agent. Here, we establish a framework that permits longitudinal analyses of the development of AI systems comparable to the development of human moral cognition. We can compare the emergence of different capacities over the course of training to the emergence of developmental milestones over the course of human development. Where the scaffolding of capacities in AI systems and human developmental trajectories cohere, this provides further confidence that the AI system has acquired the moral capacity and that the test tasks capture the moral capacity as intended (for more detail on a technical implementation of such longitudinal testing, see Section 7.3). In this way, the decomposition approach can further help cross-validate task designs for AI. This approach of decomposing a capacity and thus allowing for more rigorous cross-validation and testing could be applied to AI analyses beyond moral development (see Section 7.5).

7.2 Iterative design

Developmental psychology not only aims to describe and explain intraindividual and interindividual differences in behaviour across time, but also is fundamentally engaged in modifying or even optimising such development (Baltes et al., 2013). As Baltes et al. (2013, p. 10) contend, “the remarkable strength of a developmental approach [...] lies in its potential for preventative action and optimization. Knowledge about the history of [a] dysfunctional behaviour or problem permits interventions that direct development into more appropriate channels.” By extension, this normatively-oriented feature of developmental psychology is particularly useful in guiding the optimisation efforts that are part and parcel of a standard, iterative analysis and design approach in artificial systems.

Using this lens, our approach to moral cognitive analysis can underpin iterative design toward increasingly moral artificial agents, forming one half of a virtuous circle whereby developmental psychology advances AI research. Identifying which underlying capacities are relevant for the learning of a moral capacity can inform a training curriculum for agents. It may be the case that an underlying capacity is required (or at least is conducive) to learn a moral capacity. We can use analysis tasks to test for the influences of underlying capacities. If it was found, for example, that AI systems which previously learned joint attention more often continue to successfully learn non-rewarded helping, this suggests that this may be an underlying capacity conducive to AI systems learning helping behaviour. Similarly, if joint attention appeared dissociated from helping behaviour in AI systems, this would instead suggest that joint attention is not a required part of a training curriculum for learning to help. In one recent example of this virtuous circle, researchers found that the underwriting capacity of object-centric representation and computation was conducive to an AI system learning intuitive physical reasoning. This demonstrated the value of adding training tasks for this capacity to an overall curriculum to train AI systems in intuitive physical reasoning (Piloto et al., 2022). AI researchers may build on insights from

analyses by purposefully introducing relevant underlying capacities into a training curriculum for artificial moral cognition.

Further, analysis can be used to compare systems and reveal differences that affect the learning of artificial moral cognition. Comparing the performance of two AI systems in moral cognition tasks can reveal the reliability, robustness, and the extent to which each system displays moral capacities or underlying capacities. This may reveal that seemingly orthogonal design decisions such as AI system architecture, technical implementation, or training curriculum can influence artificial moral cognition. These implications may only become apparent upon comparative testing of two different AI systems on the same moral cognition tasks, analogous to comparative psychology studies. Findings from such comparisons can then be used to select design features that advantage artificial moral cognition. Note that although we specifically tailor the analysis approach to moral capacities in AI systems, the approach to analysis through decomposition can be applied to a range of cognitive capacities.

Lastly, as noted in Section 2, moral analysis tasks can be used to provide safety assurances for AI systems and inform iterative design on when a system is ‘safe enough’ for use. Moral tasks results can indicate to other developers or third parties which domains the AI system has been tested in, within which boundaries it is reliable and safe to use, and which performance thresholds it meets. Moreover, it is likely that thresholds of acceptable mistakes differ between capacities: for some capacities (e.g., helping a person), it may be acceptable for the AI system to fail in some contexts, whereas for other capacities (e.g., not physically harming a person), failure may be deemed unacceptable. Higher robustness and reliability thresholds may apply. Setting thresholds of when an AI system displays sufficient moral capacities for real-world use is a normative exercise and is beyond the scope of this paper.

7.3 Implications for psychology research

So far, we have discussed the merits of drawing on developmental psychology research to inform analysis of moral cognition in AI. The AI research proposed here has the potential to inform developmental psychology research — closing the virtuous circle of mutual reinforcement between AI research and developmental psychology (see also Hassabis et al., 2017). For example, by analysing moral cognition in AI systems, we may find that certain capacities thought to be interdependent in humans are dissociable in AI systems. This finding may prompt developmental psychologists to revisit these mechanisms in humans.

In this way, artificial moral cognition can function as a computational model of human moral cognition and development (Cangelosi & Schlesinger, 2018; Mareschal & Thomas, 2007; Shultz, 2003). Through this lens, we can treat current models of moral capacities in humans and their relationships as working hypotheses to test in AI systems (Bechtel & McCauley, 2020). We may also identify novel capacities that have not yet been isolated or discovered in humans. For example, there may be a capacity in AI systems similar to non-rewarded helping, and yet this may differ in key respects from how non-rewarded helping manifests in children and adults. Additionally, insights from analysing artificial moral cognition could inform our understanding of mechanisms within human moral psychology (e.g., in being able to ‘turn off’ cognitive capacities in artificial systems in ways that are impossible to do within a human sample).

AI analysis can also be seen as a way to model longitudinal accounts of moral cognitive development. One recent approach to investigating patterns in AI behaviour is to study the behaviour over the course of training time (Seidenberg & McClelland, 1989), analogous to longitudinal studies in developmental psychology (Overton, 1998). In this approach, multiple copies of the AI system are branched off at different timepoints during training (Köster et al., 2022). For example, if an AI system is trained over millions of rounds in a reinforcement learning setup, copies of the AI system can be saved prior to

training, after a certain number of rounds (e.g., 1 million rounds), and at the end of training. By comparing the moral capacities of these copies of the same AI system, analysts can study at what time point during training a particular moral capacity came online. It may be that a learning sequence or development trajectory in an AI system mirrors human developmental trajectories. The approach of detecting patterns in AI behaviour over the course of training time may generate new hypotheses on how different capacities inter-relate, and in what order they tend to come online.

We consider the analysis of moral cognition in AI systems and research into human moral development as mutually evolving and advancing. We embrace an approach of epistemic iteration (Chang, 2017), such that both the concepts of what constitutes a moral capacity (including its relationships to other antecedent capacities), as well as measurement tools to analyse this capacity, are open to revision as science progresses. Applied to our case, this means that analysis tasks measuring artificial moral cognition are open to revision as we learn about human moral development. Likewise, our understanding of human moral cognition may shift in parallel with breakthroughs in developing artificial moral cognition.

7.4 Limitations

Of course, the framework provided here is not intended to be comprehensive or sufficient for analysing all aspects of artificial moral cognition. In particular, we have focused on the foundations of moral cognition which develop in the first three to twenty-four months of human life. We contend that this is appropriate given the current state of AI research, where state-of-the-art systems do not reliably or robustly display even these basic moral capacities. We also recognise that as AI research progresses, additional tests of more sophisticated and culturally-specific moral capacities will be needed.

In a related consideration, we measure moral cognition via behavioural measurement. Some aspects of moral cognition and reasoning may be difficult to observe behaviourally.

The framework presented here is thus intended as a starting point: much like in human moral psychology, it is a first step to characterising behavioural patterns, before we can meaningfully study the underlying computational mechanisms and processes that give rise to these behaviours (Krakauer et al., 2017; Niv, 2021). Future research may expand the analysis framework by additionally translating test paradigms from neuroscientific studies of moral reasoning, such as fMRI analyses, into methods for the study of moral cognition in AI systems, such as saliency maps.

Finally, it is important to recall that our framework is built for analysing artificial moral cognition. We do not develop an approach to training or building moral cognition. Our proposal regards analysing artificial moral cognition, irrespective of how AI systems were trained. Indeed, multiple paths toward value-aligned AI systems are currently underway. One possible approach to achieving artificial moral cognition is to design a staged curriculum that roughly tracks human development. AI researchers have highlighted the benefits of such an approach: “it may not be possible to develop human-like abstract representations without the kind of developmental trajectory that human infants experience” (Mitchell, 2020). However, it is also possible that a more hands-off approach such as exposing AI systems to social learning in long-lasting environments may give rise to forms of moral cognition (Köster et al., 2022). The analysis framework proposed here can work in either scenario: it reveals to what extent these AI systems exhibit behaviour that is indicative of moral learning and moral cognition, irrespective of training.

7.5 Directions for Future Research

Future research may apply our framework to evaluate AI systems that are thought to have acquired moral capacities. Putting this framework into practice can stress-test our proposal and provide novel evidence on which moral capacities AI systems hold. This can also resolve a tension whereby psychological terms including attributions of moral or underlying capacities may be applied to AI systems without sufficient evidence (Shevlin &

Halina, 2019).

Future research may also extend our decomposition approach to alternative models of the emergence of human morality. Such approaches could rely on accounts of human evolution (Graham et al., 2013; Haidt & Joseph, 2004), specifically accounts of human cultural evolution (Boyd & Richerson, 2005), or comparative studies between species (de Waal & Aureli, 1996; Tomasello, 2019). The decomposition approach can be used to generate a framework of the chronology and interdependencies of different evolutionary milestones underpinning the emergence of moral cognitive capacities. Analogous to our approach of decomposing capacities based on human development, such an approach can test moral capacities based on their evolutionary history.

The decomposition approach introduced here could also be applied to functional cluster analyses of moral capacities, as evident in cognitive psychology. In these analyses, moral capacities are decomposed into clusters of functionally-related psychological capacities and processes. For example, moral reasoning has been decomposed into different modes of decision-making (Crockett, 2013; Moll et al., 2005). Modelling these clusters in a decomposed framework can underpin analysis of different building blocks, functionally-related capacities, and processes in artificial systems.

Finally, our approach to artificial moral cognition draws on the human example, rather than on studies of ethical behaviour by animals, human populations, businesses or institutions. To further inform potential analysis tools for AI, future work may draw on ‘moral cognition’ by these actors or systems (in addition to human cognition). Nevertheless, in our view, there are strong reasons to start by using human moral cognition as a guide for the design of AI. First, it is not clear that any of these other actors are truly capable of moral cognition. Humans may be the only guide we have for moral decision-making in the world. Second, using the human analogy as a roadmap to defining artificial moral cognition helps ensure that we understand the AI systems we build, as we can draw on the rich

history and literature on human moral development. Finally, as AI systems and humans are expected to interact, it may be beneficial to have shared moral foundations.

8. Conclusion

In this paper, we provide a framework for analysing and decomposing the complex phenomenon of moral cognition in AI systems. This allows for the evaluation and explainability of artificial moral cognition. The mode of decomposition chosen in this paper draws on developmental accounts of human moral psychology, providing a principled approach to analysing and evaluating moral capacities in artificial systems. In addition, our framework provides a novel way to test the validity of tasks used to measure artificial moral cognition. Finally, our framework sheds light on the structure of artificial moral cognition, such as interdependencies between different capacities. Advances in artificial moral cognition may then feed back into our understanding of human moral development, highlighting potential differences between human and artificial moral cognition.

Acknowledgements

We thank Patrick Butlin, Matt Crosby, Iason Gabriel, Emily Gerdin, William Isaac, Sean Legassick, Joshua May, Luis Piloto, Neil Rabinowitz, and MH Tessler for helpful comments on this project and manuscript, and Jayd Matyas for support on the design of the helping task.

References

Abramson, J., Ahuja, A., Carnevale, F., Georgiev, P., Goldin, A., Hung, A., Landon, J., Lillicrap, T., Muldal, A., Richards, B., Santoro, A., von Glehn, T., Wayne, G., Wong, N., & Yan, C. (2022). Evaluating Multimodal Interactive Agents.
<http://arxiv.org/abs/2205.13274>

- Adamson, G., Havens, J. C., & Chatila, R. (2019). Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems. *Proceedings of the IEEE*, 107(3), 518–525. <https://doi.org/10.1109/JPROC.2018.2884923>
- Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, 7(3), 149–155. <https://doi.org/10.1007/s10676-006-0004-4>
- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), 251–261. <https://doi.org/10.1080/09528130050111428>
- Anderson, M., Anderson, S., & Armen, C. (2005). Towards machine ethics: Implementing two action-based ethical theories. *Proceedings of the AAAI 2005 fall symposium on machine ethics*, 1–7.
- Anderson, M., Anderson, S. L., & Armen, C. (2006). Medethex: A prototype medical ethics advisor. *AAAI*, 1759–1765.
- Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (2013). *Life-span Developmental Psychology: Introduction To Research Methods*. Psychology Press. <https://doi.org/10.4324/9781315799704>
- Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and Nurture in Own-Race Face Processing. *Psychological Science*, 17(2), 159–163. <https://doi.org/10.1111/j.1467-9280.2006.01679.x>
- Barragan, R. C., & Dweck, C. S. (2014). Rethinking natural altruism: Simple reciprocal interactions trigger children’s benevolence. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 111(48), 17071–17074. <https://doi.org/10.1073/pnas.1419408111>
- Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B. A., Stich, S., von Rueden, C., Zhao, W., & Laurence, S. (2016). Small-scale societies

exhibit fundamental variation in the role of intentions in moral judgment.

Proceedings of the National Academy of Sciences, 113(17), 4688–4693.

<https://doi.org/10.1073/pnas.1522070113>

Bechtel, W., & McCauley, R. N. (2020). Heuristic identity theory (or back to the future):

The mind-body problem against the background of research strategies in cognitive

neuroscience. *Proceedings of the twenty first annual conference of the Cognitive*

Science Society, 67–72.

Beck, K. (2003). *Test-driven development: By example*. Addison-Wesley Professional.

Beizer, B. (1995). *Black-box testing: Techniques for functional testing of software and*

systems. John Wiley & Sons, Inc.

Benenson, J. F., Pascoe, J., & Radmore, N. (2007). Children’s altruistic behavior in the

dictator game. *Evolution and Human Behavior*, 28(3), 168–175.

<https://doi.org/10.1016/j.evolhumbehav.2006.10.003>

Bloom, P. (2013). *Just babies: The origins of good and evil*. Crown Publishers/Random

House.

Bloom, P., & Wynn, K. (2016). What develops in moral development? *Core knowledge and*

conceptual change, 279, 347–364.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous

vehicles. *Science*, 352(6293), 1573–1576.

<https://doi.org/10.1126/science.aaf2654>

Boyd, R., & Richerson, P. J. (2005). *The Origin and Evolution of Cultures*. Oxford

University Press.

Buon, M., Jacob, P., Margules, S., Brunet, I., Dutat, M., Cabrol, D., & Dupoux, E. (2014).

Friend or Foe? Early Social Evaluation of Human Interactions. *PLOS ONE*, 9(2),

e88612. <https://doi.org/10.1371/journal.pone.0088612>

Cangelosi, A., & Schlesinger, M. (2018). From babies to robots: The contribution of

developmental robotics to developmental psychology. *Child Development*

Perspectives, 12(3), 183–188.

<https://doi.org/https://doi.org/10.1111/cdep.12282>

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of science*, 63(4), 515–533.

Cave, S., Nyrupe, R., Vold, K., & Weller, A. (2019). Motivations and Risks of Machine Ethics. *Proceedings of the IEEE*, 107(3), 562–574.

<https://doi.org/10.1109/JPROC.2018.2865996>

Chang, H. (2017). Epistemic iteration and natural kinds: Realism and pluralism in taxonomy. *Philosophical issues in psychiatry IV: Psychiatric nosology*, 229–245.

Colombi, C., Liebal, K., Tomasello, M., Young, G., Warneken, F., & Rogers, S. J. (2009). Examining correlates of cooperation in autism: Imitation, joint attention, and understanding intentions. *Autism*, 13(2), 143–163.

Corbit, J., Callaghan, T., & Svetlova, M. (2020). Toddlers’ costly helping in three societies. *Journal of experimental child psychology*, 195, 104841.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, 17(8), 363–366.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests.

Psychological Bulletin, 52(4), 281–302. <https://doi.org/10.1037/h0040957>

Crosby, M. (2020). Building Thinking Machines by Solving Animal Cognition Tasks. *Minds and Machines*, 30(4), 589–615. <https://doi.org/10.1007/s11023-020-09535-6>

Crosby, M., Beyret, B., Shanahan, M., Hernández-Orallo, J., Cheke, L., & Halina, M. (2020). The Animal-AI Testbed and Competition. *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, 164–176. Retrieved August 10, 2022, from <https://proceedings.mlr.press/v123/crosby20a.html>

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, 27(1), 111–133.

[https://doi.org/10.1016/S0364-0213\(02\)00112-X](https://doi.org/10.1016/S0364-0213(02)00112-X)

- Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, 127(1), 6–21.
<https://doi.org/10.1016/j.cognition.2012.11.008>
- de Haan, M., Johnson, M. H., Maurer, D., & Perrett, D. I. (2001). Recognition of individual faces and average face prototypes by 1- and 3-month-old infants. *Cognitive Development*, 16(2), 659–678. [https://doi.org/10.1016/S0885-2014\(01\)00051-X](https://doi.org/10.1016/S0885-2014(01)00051-X)
- Dehghani, M., Tomai, E., Forbus, K. D., & Klenk, M. (2008). An integrated reasoning approach to moral decision-making. *AAAI*, 1280–1286.
- Déletang, G., Grau-Moya, J., Martic, M., Genewein, T., McGrath, T., Mikulik, V., Kunesch, M., Legg, S., & Ortega, P. A. (2021). Causal Analysis of Agent Behavior for AI Safety. Retrieved August 10, 2022, from <http://arxiv.org/abs/2103.03938>
- Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77, 1–14.
<https://doi.org/10.1016/j.robot.2015.11.012>
- de Waal, F. B. M., & Aureli, F. (1996). Consolation, reconciliation, and a possible cognitive difference between macaques and chimpanzees. *Reaching into thought: The minds of the great apes* (pp. 80–110). Cambridge University Press.
- Dietrich, F., & List, C. (2017). What matters and how it matters: A choice-theoretic representation of moral theories. *Philosophical Review*, 126(4), 421–479.
- Etel, E., & Slaughter, V. (2019). Theory of mind and peer cooperation in two play contexts. *Journal of Applied Developmental Psychology*, 60, 87–95.
- Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14), 9602–9605. <https://doi.org/10.1073/pnas.152159999>
- Fawcett, C., & Liszkowski, U. (2012). Infants anticipate others’ social preferences. *Infant and Child Development*, 21(3), 239–249.

- Gabriel, I., & Ghazavi, V. (2021). The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. <https://doi.org/10.1093/oxfordhb/9780198857815.013.18>
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. <https://doi.org/10.48550/arXiv.1811.12231>
- Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56(2), 165–193. [https://doi.org/10.1016/0010-0277\(95\)00661-h](https://doi.org/10.1016/0010-0277(95)00661-h)
- Govindarajulu, N. S., Bringsjord, S., & Peveler, M. (2019). On quantified modal theorem proving for modeling ethics. *arXiv preprint arXiv:1912.12959*.
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory. *Advances in Experimental Social Psychology* (pp. 55–130). Elsevier. <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>
- Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. <https://doi.org/10.1109/ICASSP.2013.6638947>
- Greenspan, S., & Shanker, S. (2007). The developmental pathways leading to pattern recognition, joint attention, language and cognition. *New Ideas in Psychology*, 25(2), 128–142.
- Haas, J. (2020). Moral gridworlds: A theoretical proposal for modeling artificial moral cognition. *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science*, 30(2), 219–246. <https://doi.org/10.1007/s11023-020-09524-9>
- Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4), 55–66.

- Hamlin, J. K. (2015). The case for social evaluation in preverbal infants: Gazing toward one's goal drives infants' preferences for Helpers over Hinderers in the hill paradigm. *Frontiers in Psychology*, 5. Retrieved August 10, 2022, from <https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01563>
- Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, 22(3), 186–193.
- Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad: Infants prefer those who harm dissimilar others. *Psychological Science*, 24(4), 589–594. <https://doi.org/10.1177/0956797612457785>
- Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive development*, 26(1), 30–39.
- Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450(7169), 557–559.
- Hamlin, J. K., Wynn, K., & Bloom, P. (2010). 3-month-olds show a negativity bias in their social evaluations. *Developmental science*, 13(6), 923–929. <https://doi.org/10.1111/j.1467-7687.2010.00951.x>
- Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2), 245–258. <https://doi.org/10.1016/j.neuron.2017.06.011>
- Hernández-Orallo, J. (2017). *The measure of all minds: Evaluating natural and artificial intelligence*. Cambridge University Press. <https://doi.org/10.1017/9781316594179>
- Himmelreich, J. (2018). Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethical Theory and Moral Practice*, 21(3), 669–684. <https://doi.org/10.1007/s10677-018-9896-4>

- Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, 9(2), 131–134.
<https://doi.org/10.1111/1467-9280.00024>
- Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2020). Meta-Learning in Neural Networks: A Survey. Retrieved August 10, 2022, from
<http://arxiv.org/abs/2004.05439>
- Howard, D., & Muntean, I. (2017). Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency. https://doi.org/10.1007/978-3-319-61043-6_7
- Huber, T., Weitz, K., André, E., & Amir, O. (2021). Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, 301, 103571. <https://doi.org/10.1016/j.artint.2021.103571>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
<https://doi.org/10.1038/s42256-019-0088-2>
- Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., & Irving, G. (2021). Alignment of Language Agents. Retrieved August 10, 2022, from
<http://arxiv.org/abs/2103.14659>
- Kim, T. W., Hooker, J., & Donaldson, T. (2021). Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence. *Journal of Artificial Intelligence Research*, 70, 871–890. <https://doi.org/10.1613/jair.1.12481>
- Kinzler, K. D., Shutts, K., & Correll, J. (2010). Priorities in social categories. *European Journal of Social Psychology*, 40(4), 581–592.
- Köster, R., Hadfield-Menell, D., Everett, R., Weidinger, L., Hadfield, G. K., & Leibo, J. Z. (2022). Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, 119(3), e2106028118. <https://doi.org/10.1073/pnas.2106028118>

- Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3), 480–490.
- Lancy, D. F. (2020). *Child Helpers: A Multidisciplinary Perspective* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781108769204>
- LeCun, Y. (2012). Learning invariant feature hierarchies. *European conference on computer vision*, 496–505.
- Leibo, J. Z., Dueñez-Guzman, E. A., Vezhnevets, A., Agapiou, J. P., Sunehag, P., Koster, R., Matyas, J., Beattie, C., Mordatch, I., & Graepel, T. (2021). Scalable evaluation of multi-agent reinforcement learning with melting pot. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (pp. 6187–6199). PMLR. <https://proceedings.mlr.press/v139/leibo21a.html>
- Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2017). AI Safety Gridworlds. <https://doi.org/10.48550/arXiv.1711.09883>
- Lucca, K., Capelier-Mourguy, A., Cirelli, L., Byers-Heinlein, K., Ben, R. D., Frank, M. C., Henderson, A. M. E., Kominsky, J. F., Liberman, Z., Margoni, F., Reschke, P. J., Schlingloff, L., Scott, K., Soderstrom, M., Sommerville, J., Su, Y., Tatone, D., Uzevovsky, F., Wang, Y., ... Hamlin, K. (2021). Infants' Social Evaluation of Helpers and Hinderers: A Large-Scale, Multi-Lab, Coordinated Replication Study. <https://doi.org/10.31234/osf.io/qhxkm>
- Mahajan, N., & Wynn, K. (2012). Origins of "Us" versus "Them": Prelinguistic infants prefer similar others. *Cognition*, 124(2), 227–233. <https://doi.org/10.1016/j.cognition.2012.05.003>
- Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

- 1074 Marda, V., & Narayan, S. (2020). Data in new delhi's predictive policing system.
 1075 *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*,
 1076 317–324. <https://doi.org/10.1145/3351095.3372865>
- 1077 Mareschal, D., & Thomas, M. S. (2007). Computational modeling in developmental
 1078 psychology. *IEEE Transactions on Evolutionary Computation*, 11(2), 137–150.
- 1079 McGregor, S. (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents:
 1080 The AI Incident Database. *Proceedings of the AAAI Conference on Artificial*
 1081 *Intelligence*, 35(17), 15458–15463. Retrieved August 10, 2022, from
 1082 <https://ojs.aaai.org/index.php/AAAI/article/view/17817>
- 1083 Millar, J., Lin, P., Abney, K., & Bekey, G. (2017). Ethics settings for autonomous vehicles.
 1084 *Robot ethics 2.0: From autonomous cars to artificial intelligence*, 20–34.
- 1085 Mitchell, M. (2020). On Crashing the Barrier of Meaning in Artificial Intelligence. *AI*
 1086 *Magazine*, 41(2), 86–92. <https://doi.org/10.1609/aimag.v41i2.5259>
- 1087 Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial Theory as
 1088 Sociotechnical Foresight in Artificial Intelligence. *Philosophy and Technology*, 33(4),
 1089 659–684. <https://doi.org/10.1007/s13347-020-00405-8>
- 1090 Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural
 1091 basis of human moral cognition. *Nature Reviews Neuroscience*, 6(10), 799–809.
 1092 <https://doi.org/10.1038/nrn1768>
- 1093 Moore, C., Dunham, P. J., & Dunham, P. (2014). *Joint attention: Its origins and role in*
 1094 *development*. Psychology Press.
- 1095 Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral*
 1096 *Neuroscience*, 135(5), 601. <https://doi.org/10.1037/bne0000471>
- 1097 Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars:
 1098 An Applied Trolley Problem? *Ethical Theory and Moral Practice*, 19(5), 1275–1289.
 1099 <https://doi.org/10.1007/s10677-016-9745-2>

- 1100 Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?
 1101 *science*, 308(5719), 255–258.
- 1102 Opfer, J. E., & Gelman, S. A. (2011). Development of the animate-inanimate distinction.
 1103 *The Wiley-Blackwell handbook of childhood cognitive development*, 2nd ed.
 1104 (pp. 213–238). Wiley-Blackwell.
- 1105 Ortega, W., Maini, V., & the DeepMind Safety Team. (2018). Building safe artificial
 1106 intelligence: Specification, robustness, and assurance. *Medium*. Retrieved August 10,
 1107 2022, from [https://deepmindsafetyresearch.medium.com/building-safe](https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1)
 1108 [-artificial-intelligence-52f5f75058f1](https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1)
- 1109 Over, H., & Carpenter, M. (2009). Eighteen-month-old infants show increased helping
 1110 following priming with affiliation. *Psychological Science*, 20(10), 1189–1193.
- 1111 Overton, W. (1998). Developmental psychology: Philosophy, concepts, and methodology.
 1112 *Theoretical Models of Human Development, Vol. 1: Handbook of Child Psychology*
 1113 (pp. 107–188).
- 1114 Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics
 1115 learning in a deep-learning model inspired by developmental psychology. *Nature*
 1116 *Human Behaviour*. <https://doi.org/10.1038/s41562-022-01394-8>
- 1117 Prinz, J. (2008). Is morality innate. *Moral psychology*, 1, 367–406.
- 1118 Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B.,
 1119 Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap:
 1120 Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of*
 1121 *the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.
 1122 <https://doi.org/10.1145/3351095.3372873>
- 1123 Rautenbach, G., & Keet, C. M. (2020). Toward equipping Artificial Moral Agents with
 1124 multiple ethical theories. <https://doi.org/10.48550/arXiv.2003.00935>
- 1125 Rodriguez-Soto, M., Serramia, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2022).
 1126 Instilling moral value alignment by means of multi-objective reinforcement learning.

Ethics and Information Technology, 24(1), 9.

<https://doi.org/10.1007/s10676-022-09635-0>

Roff, H. M. (2020). Expected Utilitarianism.

<https://doi.org/10.48550/arXiv.2008.07321>

Sagi, A., & Hoffman, M. L. (1976). Empathic distress in the newborn. *Developmental Psychology*, 12(2), 175–176. <https://doi.org/10.1037/0012-1649.12.2.175>

Santoni de Sio, F. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice*, 20(2), 411–429.

Sanz, R. (2020). ETHICA EX MACHINA. Exploring artificial moral agency or the possibility of computable ethics. *Zeitschrift für Ethik und Moralphilosophie*, 3(2), 223–239. <https://doi.org/10.1007/s42048-020-00064-6>

Scarf, D., Imuta, K., Colombo, M., & Hayne, H. (2012). Social evaluation or simple association? Simple associations may explain moral reasoning in infants. *PloS One*, 7(8), e42698. <https://doi.org/10.1371/journal.pone.0042698>

Scola, C., Holvoet, C., Arciszewski, T., & Picard, D. (2015). Further Evidence for Infants' Preference for Prosocial Over Antisocial Behaviors. *Infancy*, 20(6), 684–692. <https://doi.org/10.1111/infa.12095>

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4), 523.

Shanahan, M., Crosby, M., Beyret, B., & Cheke, L. (2020). Artificial intelligence and the common sense of animals. *Trends in Cognitive Sciences*, 24(11), 862–872. <https://doi.org/https://doi.org/10.1016/j.tics.2020.09.002>

Sheskin, M., Bloom, P., & Wynn, K. (2014). Anti-equality: Social comparison in young children. *Cognition*, 130(2), 152–156.

Shevlin, H., & Halina, M. (2019). Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, 1(4), 165–167. <https://doi.org/10.1038/s42256-019-0039-y>

- 1154 Shultz, T. R. (2003). *Computational developmental psychology*. Mit Press.
- 1155 Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G.,
1156 Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S.,
1157 Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M.,
1158 Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go
1159 with deep neural networks and tree search. *Nature*, 529(7587), 484–489.
1160 <https://doi.org/10.1038/nature16961>
- 1161 Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A.,
1162 Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L.,
1163 van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of
1164 Go without human knowledge. *Nature*, 550(7676), 354–359.
1165 <https://doi.org/10.1038/nature24270>
- 1166 Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial*
1167 *Intelligence*, 299, 103535.
- 1168 Singer, D. G., & Singer, J. L. (2009). *Imagination and Play in the Electronic Age*. Harvard
1169 University Press.
- 1170 Singh, L. (2022). Automated Kantian Ethics. *Proceedings of the 2022 AAAI/ACM*
1171 *Conference on AI, Ethics, and Society*, 915.
1172 <https://doi.org/10.1145/3514094.3539527>
- 1173 Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness?
1174 *Psychological Science*, 23(2), 196–204.
1175 <https://doi.org/10.1177/0956797611422072>
- 1176 Smith, G. T. (2005). On Construct Validity: Issues of Method and Measurement.
1177 *Psychological Assessment*, 17(4), 396–408.
1178 <https://doi.org/10.1037/1040-3590.17.4.396>

- 1179 Spelke, E., Lee, S. A., & Izard, V. (2010). Beyond core knowledge: Natural geometry.
 1180 *Cognitive Science*, 34(5), 863–884.
 1181 <https://doi.org/10.1111/j.1551-6709.2010.01110.x>
- 1182 Spelke, E. S. (2003). What makes us smart? Core knowledge and natural language.
 1183 *Language in mind: Advances in the study of language and thought* (pp. 277–311).
 1184 MIT Press.
- 1185 Spelke, E. S. (2016). Core knowledge and conceptual change. *Core knowledge and*
 1186 *conceptual change*, 279, 279–300.
- 1187 Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, 10(1),
 1188 89–96.
- 1189 Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies.
 1190 *Nature Human Behaviour*, 1(4), 1–7.
- 1191 Sterelny, K., & Fraser, B. (2017). Evolution and Moral Realism. *The British Journal for the*
 1192 *Philosophy of Science*, 68(4), 981–1006. <https://doi.org/10.1093/bjps/axv060>
- 1193 Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and
 1194 methodology. *Annual review of clinical psychology*, 5, 1.
- 1195 Tasimi, A., & Wynn, K. (2016). Costly rejection of wrongdoers by infants and children.
 1196 *Cognition*, 151, 76–79. <https://doi.org/10.1016/j.cognition.2016.03.004>
- 1197 Taylor, A. H., Bastos, A. P. M., Brown, R. L., & Allen, C. (2022). The signature-testing
 1198 approach to mapping biological and artificial intelligences. *Trends in Cognitive*
 1199 *Sciences*. <https://doi.org/10.1016/j.tics.2022.06.002>
- 1200 Teller, D. Y. (1979). The forced-choice preferential looking procedure: A psychophysical
 1201 technique for use with human infants. *Infant Behavior & Development*, 2(2),
 1202 135–153. [https://doi.org/10.1016/S0163-6383\(79\)80016-8](https://doi.org/10.1016/S0163-6383(79)80016-8)
- 1203 Ting, F., Dawkins, M. B., Stavans, M., & Baillargeon, R. (2020). Principles and concepts
 1204 in early moral cognition. *The social brain: A developmental perspective* (pp. 41–65).
 1205 The MIT Press. <https://doi.org/10.7551/mitpress/11970.001.0001>

- 1206 Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Belknap Press of Harvard
 1207 University Press. <https://doi.org/10.4159/9780674988651>
- 1208 Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in
 1209 the gaze following of great apes and human infants: The cooperative eye hypothesis.
 1210 *Journal of Human Evolution*, 52(3), 314–320.
 1211 <https://doi.org/10.1016/j.jhevol.2006.10.001>
- 1212 Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the
 1213 simulation theory of cognition. *Cognitive Systems Research*, 48, 56–66.
 1214 <https://doi.org/10.1016/j.cogsys.2017.04.002>
- 1215 Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J.,
 1216 Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M.,
 1217 Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., ...
 1218 Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement
 1219 learning. *Nature*, 575(7782), 350–354.
 1220 <https://doi.org/10.1038/s41586-019-1724-z>
- 1221 Wallach, W., & Marchant, G. (2019). Toward the Agile and Comprehensive International
 1222 Governance of AI and Robotics [point of view]. *Proceedings of the IEEE*, 107(3),
 1223 505–508. <https://doi.org/10.1109/JPROC.2019.2899422>
- 1224 Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young
 1225 chimpanzees. *Science (New York, N.Y.)*, 311(5765), 1301–1303.
 1226 <https://doi.org/10.1126/science.1121448>
- 1227 Warneken, F., & Tomasello, M. (2007). Helping and Cooperation at 14 Months of Age.
 1228 *Infancy*, 11(3), 271–294.
 1229 <https://doi.org/10.1111/j.1532-7078.2007.tb00227.x>
- 1230 Warneken, F., & Tomasello, M. (2009a). The roots of human altruism. *British Journal of*
 1231 *Psychology*, 100(3), 455–471. <https://doi.org/10.1348/000712608X379061>

- 1232 Warneken, F., & Tomasello, M. (2009b). The roots of human altruism. *British Journal of*
1233 *Psychology*, 100(3), 455–471.
- 1234 Warneken, F., & Tomasello, M. (2009c). Varieties of altruism in children and chimpanzees.
1235 *Trends in cognitive sciences*, 13(9), 397–402.
- 1236 Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining
1237 function of wrong beliefs in young children's understanding of deception. *Cognition*,
1238 13(1), 103–128.
- 1239 Wittig, M., Jensen, K., & Tomasello, M. (2013). Five-year-olds understand fair as equal in
1240 a mini-ultimatum game. *Journal of experimental child psychology*, 116(2), 324–337.
- 1241 Woo, B. M., Tan, E., & Hamlin, K. (2022). Human morality is based on an early-emerging
1242 moral core. <https://doi.org/10.31234/osf.io/98d36>
- 1243 Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach.
1244 *Cognition*, 69(1), 1–34.
- 1245 Wu, Z., Pan, J., Su, Y., & Gros-Louis, J. (2013). How joint attention relates to cooperation
1246 in 1- and 2-year-olds. *International Journal of Behavioral Development*, 37(6),
1247 542–548. <https://doi.org/10.1177/0165025413505264>
- 1248 Wynn, K. (2016). Origins of Value Conflict: Babies Do Not Agree to Disagree. *Trends in*
1249 *Cognitive Sciences*, 20(1), 3–5. <https://doi.org/10.1016/j.tics.2015.08.018>
- 1250 Wynn, K., Bloom, P., Jordan, A., Marshall, J., & Sheskin, M. (2018). Not noble savages
1251 after all: Limits to early altruism. *Current Directions in Psychological Science*,
1252 27(1), 3–8. <https://doi.org/10.1177/0963721417734875>