AI and Ethics (2025) 5:2071–2087
https://doi.org/10.1007/s43681-024-00536-0

ORIGINAL RESEARCH

Minimum levels of interpretability for artificial moral agents

Avish Vijayaraghavan[1,2,3] · **Cosmin Badea**[3]

Received: 6 September 2023 / Accepted: 21 July 2024 / Published online: 31 July 2024
© The Author(s) 2024

Abstract
As artificial intelligence (AI) models continue to scale up, they are becoming more capable and integrated into various forms of decision-making systems. For models involved in moral decision-making (MDM), also known as artificial moral agents (AMA), interpretability provides a way to trust and understand the agent's internal reasoning mechanisms for effective use and error correction. In this paper, we bridge the technical approaches to interpretability with construction of AMAs to establish minimal safety requirements for deployed AMAs. We begin by providing an overview of AI interpretability in the context of MDM, thereby framing different levels of interpretability (or transparency) in relation to the different ways of constructing AMAs. Introducing the concept of the Minimum Level of Interpretability (MLI) and drawing on examples from the field, we explore two overarching questions: whether a lack of model transparency prevents trust and whether model transparency helps us sufficiently understand AMAs. Finally, we conclude by recommending specific MLIs for various types of agent constructions, aiming to facilitate their safe deployment in real-world scenarios.

Keywords Interpretability · Artificial intelligence · Artificial moral agents · Moral decision-making

Abbreviations

AI	Artificial intelligence
AMA	Artificial moral agent
BU	Bottom-up
GPT	Generative pre-trained transformer
IML	Interpretable machine learning (or interpretability)
LLM	Large language model
MDM	Moral decision-making
ML	Machine learning
MLI	Minimum level of interpretability
TD	Top-down

1 Introduction

The deployment of consumer-facing generative artificial intelligence (AI) models such as Midjourney and ChatGPT has raised important questions on the ethics [1] and consequences of widespread access to AI technologies [2]. Tracing the evolution of these models over the past five years [3], it is likely that we will soon see multi-modal general-purpose models [4–8] available to the public. As these models begin operating with higher autonomy and become integrated into existing applications [9–11] (e.g. ChatGPT with plugins, AI vision models within self-driving cars), they will play a greater role in many aspects of human decision-making [12, 13]. A fundamental subset of human decision-making is moral decision-making (MDM). MDM comes in many forms—some examples include predicting whether criminals will reoffend [14], deciding appropriate treatment plans for patients [15], and executing military defense strategies in a way that is compliant with original mission orders [16]. MDM is difficult because it often involves weighing competing values in complex and ambiguous situations [17]. Where other types of decision-making may be based in pragmatic considerations, like efficiency or performance, MDM requires making judgements about what is right and wrong.

AI models that are involved in MDM are called artificial moral agents (AMAs). For these types of decisions, it is

✉ Avish Vijayaraghavan
avish.vijayaraghavan17@imperial.ac.uk

[1] Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK

[2] UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, UK

[3] Department of Computing, Imperial College London, London, UK

imperative that we have high levels of agent understanding so that errors can be corrected swiftly and performance better aligned with human values to prevent unintended and potentially harmful agent effects [9, 13]. "Understanding" an agent can take on different levels of complexity [18, 19] which require different AMA constructions [13] for effective deployment. Here, effective deployment means finding a model of appropriate capacity for its task so it can be deployed and updated as efficiently as possible in the real-world. This understanding of agent behaviour is enabled through the field of interpretable machine learning (IML, a.k.a. interpretability) which helps make AI models more trustworthy and transparent [18, 20].

Taking the necessity of interpretability in AMAs as a spectrum, we frame our discussion less around whether AMAs need interpretability (i.e. a binary decision), and more around the Minimum Level of Interpretability (MLI) for different AMA constructions. As such, this paper revolves around three key concepts: **different AMA constructions**, **different levels of interpretability**, and **how AMA capability is altered by different levels of interpretability**. Much of machine ethics research concerns aligning ethical schemas with machines [21]. Our work continues down this line, aiming to bridge technical aspects of interpretability with AMA construction for smoother deployment. An important disclaimer is that this field is nascent and we are proposing general safety rules based on limited evidence—the MLI will evolve as more evidence is gathered for different AMA use cases. Our scope is limited to a computational understanding of moral decision-making in AMAs and we do not consider the problem of where responsibility falls for agent decisions [22].

2 Background and current work

2.1 Improving moral decision-making (MDM)

We adopt the definition given by Garrigan and colleagues [23] that moral decision-making (MDM) is "any decision, including judgements, evaluations, and response choices, made within the 'moral domain' " with the moral domain consisting of decisions concerning issues like harm and fairness. The authors show how theories for the development of morality in humans fall into three main categories: cognitive, affective (or emotional), and social [23] (shown in Fig. 1). Cognitive theories take on a neuroscience-based approach (i.e. which parts of the brain are activated in response to moral stimuli), affective theories are based primarily in developmental psychology, and social theories reflect how moral psychology and behaviours changes from an individual to a population [23]. We recap the authors' original summary of the various theories to illustrate how they have evolved in complexity. Theories from developmental psychology look at how moral reasoning develops, firstly in childhood [24], then beyond into adolescence and adulthood [25], and now incorporate a spectrum based in cognitive theories like perspective taking and scripts [26] as well as attention and working memory [27]. The most well-known affective theory—Haidt's social intuitionist theory [28, 29]—investigates how quicker emotional decisions are rationalised post-hoc into "moral" decisions (akin to System 1 of Kahneman's reflexive System 1 and more calculating System 2 ways of thinking [30]). Other affective theories like dual-process theory have incorporated neuroscience aspects by focusing on activity levels of parts of the brain with real-time MDM [31, 32].

MDM necessitates navigating trade-offs between different interests, such as those of individuals, groups, or society as a whole [33–35]. This makes MDM emotionally challenging since it involves choices that have significant consequences

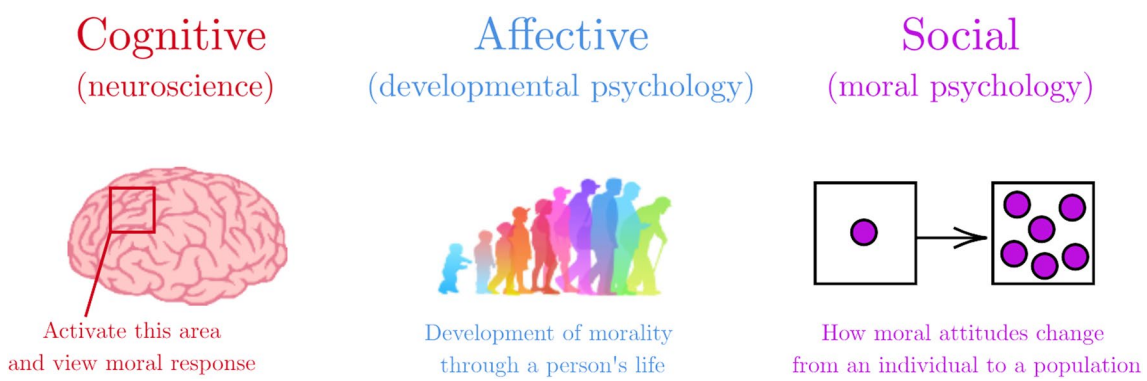


Fig. 1 The three main theories for moral decision-making

for oneself or others with positive and negative effects often getting amplified with scale [36]. The inclusion of AI models in emotional decisions may initially seem off-putting [37]. But the emotional challenge of making important decisions is the exact reason that we want non-emotional agents involved—so they can minimise human inconsistency [38] and provide fairer outcomes [18, 20]. We are careful here to avoid referencing "automation" of human decisions [39]. Real-world moral decisions have multiple reasoning steps and are challenging for two reasons: (1) they can have multiple valid (and invalid) explanations, (2) they are complex decisions built up of individual reasoning steps that do not necessarily belong to the same moral paradigm, such as with mixed strategies in game theory (e.g. your final decision could be a combination of Utilitarian reasoning and Virtue Ethics reasoning). We illustrate this concept more clearly in Fig. 2. When we move these concepts over to the world of machine learning, we fall into the interpretability challenges described in Fig. 4.

Instead, we look to improve MDM by integrating AI into standard human decision-making processes in stages which are prone to human error or with data that are beyond our cognitive capacities [40–43]. So, how do we integrate moral psychology theories into our AI models? Moral philosophies such as Deontology allow us to formalise aspects of MDM but the multi-factorial development of morality in humans is hard to represent as just one moral philosophy [44]. Context-specific models may be enabled by singular theories such as Virtue Ethics for general clinical settings [45] or Bentham's Felicific Calculus for end-of-life situations [46] but more flexible constructions are also possible. These flexible constructions help generalise MDM systems to unseen and novel moral situations [47]. We outline the overarching study of these models, called artifical moral agents, in the next section.

2.2 Different constructions of artificial moral agents (AMAs)

The study of artificial moral agents (AMAs) is an interdisciplinary field between computer science, ethics, and philosophy. As such, we first clarify terminology. The terms "model" and "agent" both refer to AI systems, and we use agent to emphasise that the model has a degree of autonomy. "Morals" concern actions of virtue and "morality" reflects that these behaviours are practised habitually to become things we accept internally and externally as rules or principles [13]—more concisely, morals "regulate selfishness and make social life possible" [17]. "Ethics" is a broader term than morality which can be defined as a "rational reflection on moral behaviours" [13] and better emphasises contextual differences for moral behaviours [48]. The words are closely linked and for our purposes can be used interchangeably, but we refer solely to morality going forward since this is the terminology of AMAs. Thus, an AMA is a program that can act or make decisions in a "moral" way, with a degree of autonomy [13]. The autonomy of an AMA is the extent to which a human can interact with the agent to change one of its decisions [9]. There are three categories of AMAs distinguished by the level of moral consideration built into them and that they can act on: implicit, explicit, and full [49]. Implicit AMAs cannot distinguish good from bad behaviour but are constructed to enable moral behaviour, explicit agents use inbuilt ethical rules (e.g. from logical formalisms or algorithmic constraints), while full ethical agents, like humans, possess aspects of consciousness like desires, intentions, and free will. We only consider AMAs within the first two levels to ignore questions related to fair treatment of potentially sentient artificial agents, limiting the scope of interpretability requirements to human (and not machine) safety. The breadth of these categories makes it more challenging to analyse how their differences manifest in real-world agents so we turn our attention to more granular parameters of AMA construction: the moral paradigm, the scale, and the purpose of the agent, all shown in Fig. 3.

We group moral philosophies and moral psychologies under the name moral paradigm or framework which tell us how morality is instilled into the agent. There are three broad moral paradigms we consider: top-down (TD), bottom-up (BU), or hybrid. TD approaches start from a set of principles or a moral framework (e.g. Utilitarianism), BU approaches have no moral framework and instead aim to learn morality from the environment, and hybrid approaches combine aspects of the two [50]. Agents, like standard AI models, can be constructed at different scales which produce different performance capabilities [51]—we consider standard

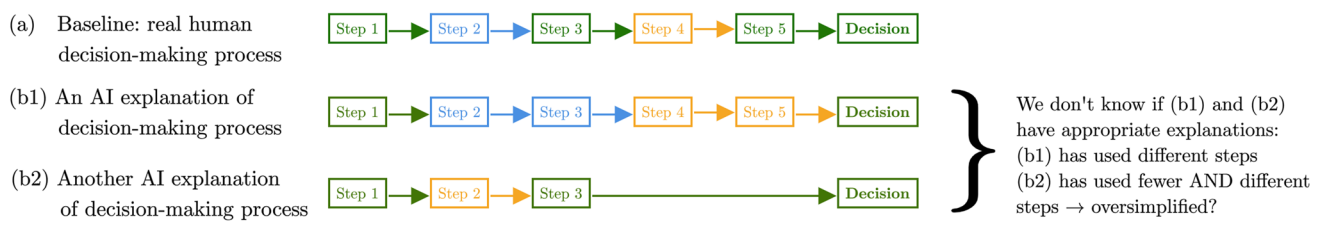


Fig. 2 The challenges of ensuring appropriate explanations for complex decisions

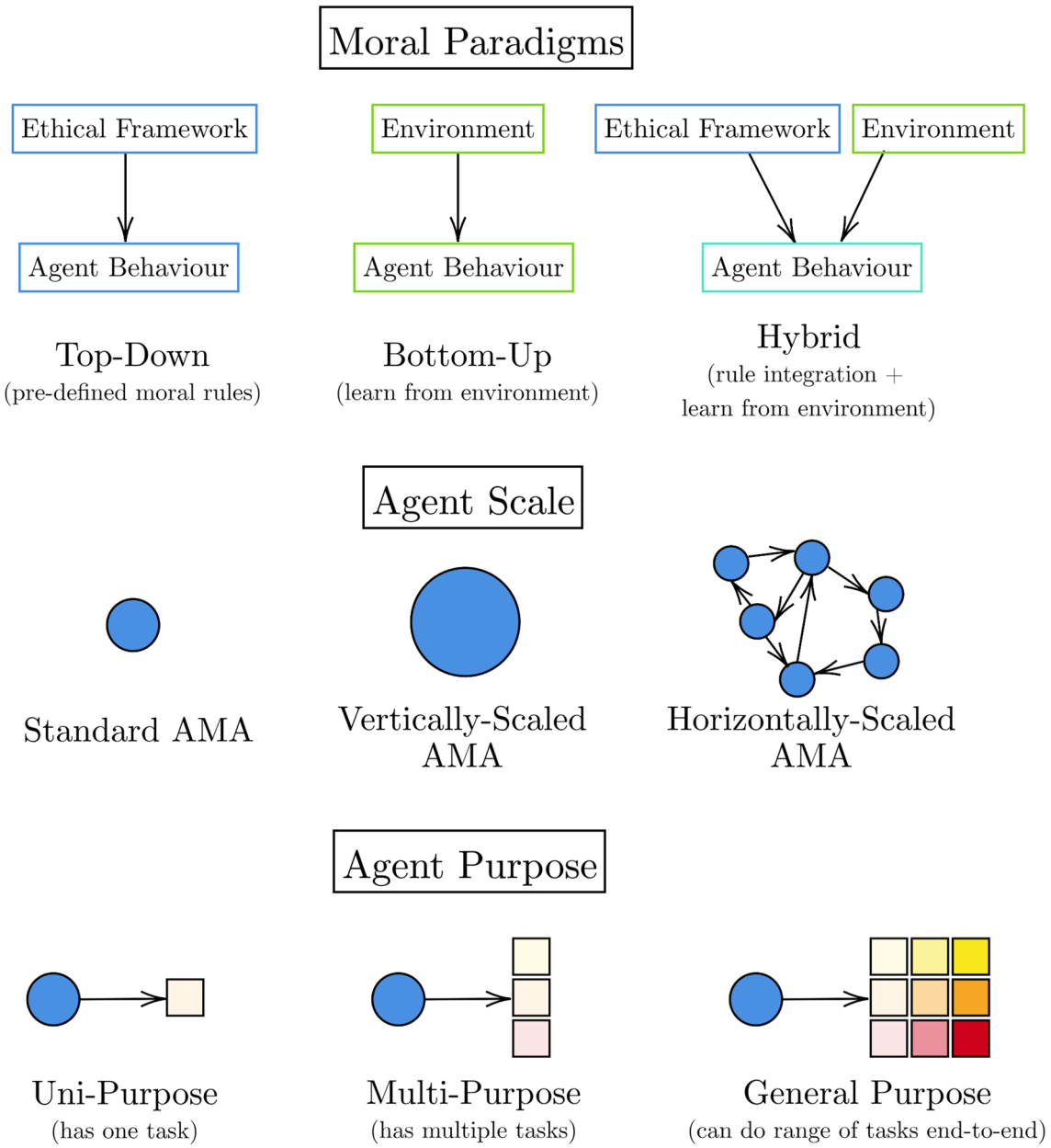


Fig. 3 Diagrams of how different agents can be constructed. Blue circles are agents (color figure online)

individual agents, high capacity individual agents (vertical scaling), and multi-agent systems (horizontal scaling) [52]. For simplicity in horizontally-scaled systems, we assume all agents are cooperative and that there are no unpredictable agent-agent interaction effects [13]. The purpose of the AMA is the task that it is designed to do and can be split into *uni-purpose, multi-purpose, and general-purpose*. The distinction between multi-purpose and general-purpose gets blurred as agents become more capable at multiple tasks and so, to avoid case-by-case analysis of different purposes, we focus on the distinction between uni-purpose and general-purpose.

2.3 Are we considering "true" moral agents?

We ask ourselves this question to clearly define the scope of this paper on the philosophical and psychological sides. We break it down into two interconnected questions: (1) how is an AMA able to perform moral decision-making (MDM) instead of just standard decision-making, and (2) is the MDM capacity of AMAs discussed here sufficient to view our AMAs as "true" moral agents?

Starting with Question (1), the capacity for decision-making leads into moral agency when the decision-making agent (human or machine) has a notion of moral responsibility

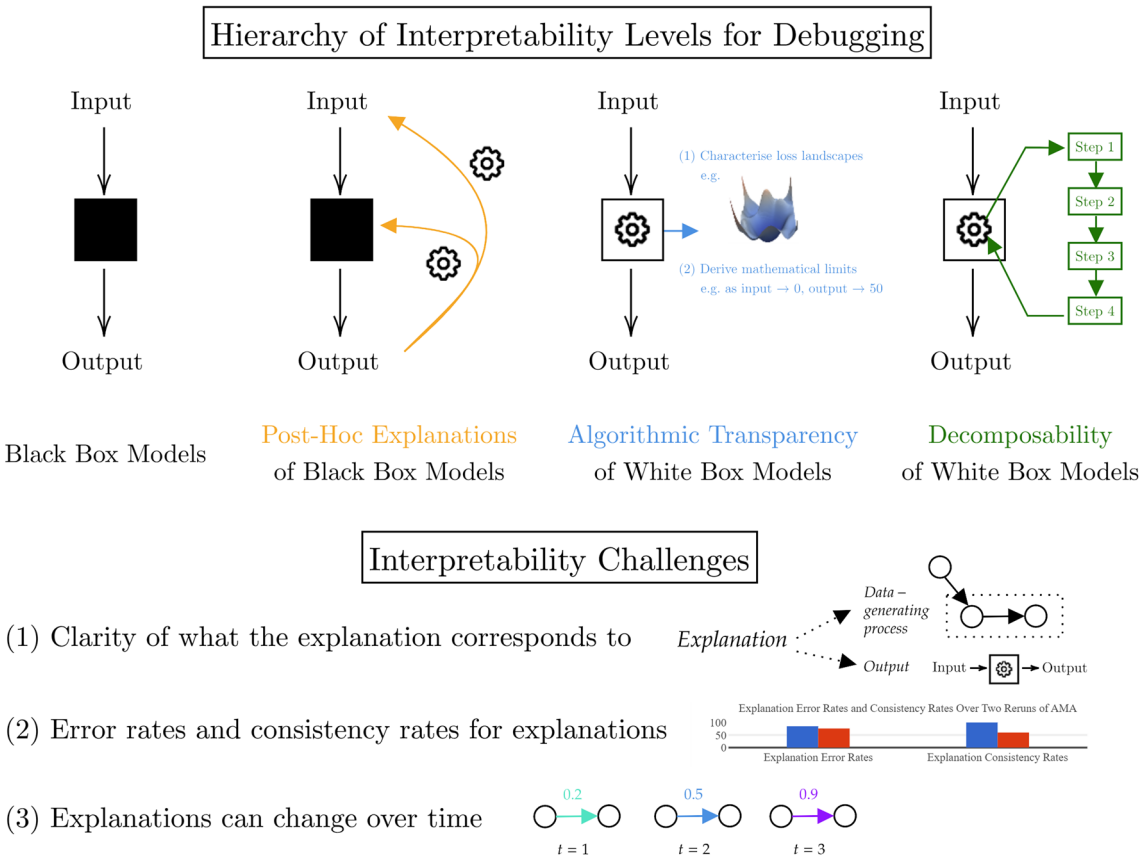


Fig. 4 Hierarchy of interpretability levels used in this work alongside the main challenges for interpretability. Black squares are black box models, white squares are white box models, the cog is the reasoning that is visible or not visible

and has learned to generalise to new MDM situations, often through interacting with the environment and responding to consequences of its actions [53]. In humans, these notions of responsibility and generalisability develop alongside each other as people's brains mature from babies to adults (as per affective theories of MDM [23]), and as humans interact more with other humans in society and develop a shared sense of understanding, trust, and ultimately responsibility for other around them (as per social theories of MDM [23]). Although current machines can be seen to operate in constrained environments, have limited sensory perception, and have restricted relationships with other machines, we can still imbue them with some sense of morality (i.e. TD or hybrid AMAs in Fig. 3) which can improve (i.e. through standard AI training techniques), and share information to other machines (like horizontal scaling in Fig. 3). In this way, although AMAs do not share the full spectrum of human experience, we can approximate moral development in machines. As mentioned in Sect. 2.1, we do not aim to automate MDM in this paper as others do [54]. Rather, we envisage a simpler approximate solution in the real world which is to give machines greater capacity for moral agency by including human-like feedback into the machine's learning process. As such, the AMAs discussed in this paper are able to make (approximate) moral decisions.

Moving to Question (2), several papers have discussed whether AMAs constitute "true" moral agents in the philosophy and ethics literature [53, 55–58]. The standard view of AMAs requires human-like consciousness as the starting point of the discussion [56]. However, we focus on an alternative perspective, called the functional view of AMAs which requires three different things [55]: interactivity (that the agent interacts with its environment), it has some level of independence (which we refer to as "autonomy"), and adaptability (that the level of independence changes based on how the agent interacts with the environment). The functional view allows us to centre the discussion around practical decision-making which is more relevant to our aims of bridging AMAs with a technical field like interpretability [58]. As mentioned in Sect. 2.1, our paper is written to integrate AI and human decision-making, and so although the AMAs we consider are learning and interacting only in simple ways with the environment (i.e. learning and interacting from limited environment-specific data), they are still interacting and learning; and although they are mostly used by humans (i.e. lose some autonomy), they still have some

autonomy themselves; and thus they still satisfy, to some degree, all three criteria, and this we see as sufficient to be considered "true" moral agents. While we do not aim to fully solve the question of true moral agency here, we focus on AMAs that can justifiably be considered moral agents based on the functional perspective, as described above, without needing to claim they have achieved all aspects of human moral agency.

2.4 Interpretability "levels" and their importance for MDM

There is no agreed-upon definition for interpretability but it can be viewed generally as a domain-specific quality for understanding or trusting our agent [18, 20]. Two seminal perspectives from the explainability/interpretability literature pose a dichotomy where "explainability" is using a black box model and then explaining it with a secondary post-hoc model, and "interpretability" is not using a black box, instead using a model that explains itself (a.k.a. a white box or transparent model) [18, 20]. Lipton [18] adds further detail based on three different paradigms of white box modelling: algorithmic transparency, decomposability, and simulatability. Algorithmic transparency amounts to a formal understanding of the agent's learning process [18], for example, better characterisation of the loss surface [59, 60] or providing internal convergence properties [61]. Decomposability corresponds to transparency at the level of model parameters while simulatability is transparency across the whole model [18]. Within the context of MDM, decomposability corresponds to having an intuitive and step-by-step explanation for each major agent decision [12]. We believe that complete decomposability (i.e. intuitive explanations for all model parameters and output decisions) subsumes simulatability and so do not consider simulatability further. For clarity, we use "transparency" when referring to forms of white box agents, "post-hoc explainability" for explanations of black box agents, and "interpretability" when referring to both of these concepts together.

Interpretability alone is not necessarily useful for our domain of MDM but becomes so when directed towards a specific goal [62]. Watson lays out three challenges that interpretability faces [19] (summarised in Fig. 4): clarity of what the explanation corresponds to (e.g. the model's outputs, the data generating process, different sub-objectives), error rates and consistency rates for explanations, and little consideration of the fact that explanations can change over time. Given these challenges are important for all interpretable systems, we take the aim of interpretability in MDM as providing an understanding of agent decision-making processes for appropriate error correction, error prevention, and agent behaviour optimisation. More simply, interpretability is useful here as a debugging tool for different stakeholders involved in a moral decision [40]. While there are clearer situations where interpretability is not needed, such as when an agent is not involved in a high-stakes decision [20] or does not have a significant impact on society [63], we assume all decisions requiring moral consideration as potentially high-stakes and enabled by reliable human-agent collaboration [42, 43, 64], and thus in need of some form of interpretability. This lends itself to our characterisation of interpretability as a spectrum more than a binary requirement.

While the different types of interpretability do not fall neatly into a hierarchy of explanation complexity [18], this becomes easier when each type is viewed as a debugging tool. Certain types of interpretability are more challenging to program into an agent and different types are required depending on what the AMA does and the number of stakeholders involved. A loose interpretability hierarchy in terms of increasing agent construction difficulty, which we phrase as "levels", is shown in Fig. 4 and is as follows: black box models—post-hoc explanations of black box models—algorithmic transparency—decomposability.

Here, we have described the different types of interpretability from an AI perspective. In the following sections, we map these interpretability levels onto moral decision-making performed by AMAs and qualitatively determine the minimum level of interpretability (MLI) required in different contexts. We start by asking ourselves whether black box AMAs allow for trust and continue with whether transparency is key to understanding all AMAs.

3 Does lack of transparency in AMAs prevent trust?

In this section, we discuss the reliability of moral reasoning in AMAs without transparency, commonly seen as black box AMAs. Thus, we will use "black box" as a generic term for such systems. Since the internal agent reasoning is unavailable to us directly, we require a level of faith in our agent, which corresponds to framing **interpretability as trust** for MDM [18]. Trust can take on a range of meanings: confidence in the agent to make the correct decision, the consistency of the agent's decisions in certain situations, or whether the agent makes decisions that are right or wrong in a human-like manner. We discuss whether AMAs can learn moral principles, and then if so, whether these principles are appropriate. We conclude that we can trust AMAs without transparency if they adopt the benefits of both BU and TD agents.

3.1 Can we tell if black box AMAs have learned *any* moral principles?

We define the "environment" of an AMA as the potential hypothesis space spanned by the data and the AMA's

learning process, which consists of its internal model and training regime. BU AMAs are predicated on the existence of functional morality [65]—that agents are able to learn morals from their environment. With a black box BU AMA, how can we be sure that our agent has learned some form of morality? Allen et al. [66] defined a Moral Turing Test which states that "if two systems are input–output equivalent, they have the same moral status", with the subsequent debate following that of the Chinese Room Argument [67]. Effectively, any machine capable of memorising a sufficiently diverse and framework-like or human-like set of input–output moral relationships would pass the test, even though we would not be able to determine if it is intrinsically "moral".

There is no universal definition for morality in humans beyond notions of obligation [68]. However, psychological theories for the development of morality within humans [23, 69] and the "teachability" of moral values [70, 71] point to the development of morality as a process [23], if an imprecise one. It is from the process of learning via experience in the world that we as humans feel the obligation to be a moral agent [68]. The same argument has been made for AI models [72] through development of value-based agents which learn human values, a process which has been successfully implemented in the multi-valued action reasoning system [54]. The lack of a universal definition means that memorisation could suffice as a type of morality, particularly if we consider memorisation a form of learning [73]. However, it is insufficient for trusting our agent because there are no guarantees that the agent will generalise outside the training set. This is an issue that can be explained more clearly by the difference between form and meaning as expressed initially by Badea and Artus [72] and subsequently formalised by Bender and colleagues [74]. They state that if an agent is trained only on form (e.g. pixels, words, etc.) without any input for communicative intent behind the form (i.e. context-dependent meaning, which exists at different scopes of worldly experience [75]), it cannot truly intuit any meaning of morality, and certainly no definition that extends to novel contexts [23, 47, 76]. This comes down to the "The Interpretation Problem" [72], which refers to the issue of endless potential interpretations for any symbolic representation given to an AMA. This makes it impossible to guarantee a fully accurate transmission of meaning regardless of the medium chosen.

The question then arises: would sufficient memorisation of triples with a structure of (input, output, communicative intent) suffice for "learning morality"? The ever-changing nature of communicative intent across cultures and over time requires a potentially infinite and unobtainable set of such data [75] which prevents learning a cohesive set of moral principles [74] unless the context is clearly defined a priori and an acceptable level of error with an "action limit" defined [77]. To avoid the difficulty of defining a complete and ethical training dataset for an imprecise objective [47, 78–80], we can instead use black box TD AMAs to approach the problem from a different angle: ensure the agent has a set of pre-defined moral principles rather than relying on the data and where it has come from. TD AMAs use a specific moral framework (or set of frameworks) which allow us to compare the TD agent's input–output moral relationships with the most likely output from that same moral framework. While the TD construction is more precise than the BU construction, it enables greater trust in the AMA's learned principles because we have a form of ground truth that is less variable than individual agent comparisons permitted by the Moral Turing Test for BU constructions. This concept of precision is shown in Fig. 5. Hybrid settings require additional domain knowledge but are even better since they can find the right "compromise between being too flexible and too strict" [81]. From this, we deduce that morality can be learned by agents once it has some initial framework for a given domain, and further capacity can be enabled by data with communicative intents.

3.2 Can we guarantee that black box AMAs have learned *appropriate* moral principles?

Let us assume that our AMA is able to learn moral principles from its environment. We are now presented with a different problem: how can we be sure that this environment reflects our desired human values and that the agent is learning them? Current inequalities in our world have been shaped (and are still influenced) by cultural remnants of historically unequal power dynamics [82, 83]. An overt example is historical medical exploitation of underrepresented communities which has led to a lack of diversity in large-scale genomic data, and been a major obstacle to generalisable genomic insights across populations [84]. Similarly damaging, but more subtle, is the inadequate treatment of sensitive variables (e.g. age, sex, race, etc.) which can lead to models shortcutting to high predictive accuracy based on harmful stereotypes [20, 85]—textbook examples include racist explanations in criminal recidivism prediction [14] and proxy variables that consolidate racial disparities in population health models for medical support prioritisation [86]. Beyond error correction, models can also serve as ways to improve existing disparities—for example, enabling smoother socioeconomic mobility via smarter intergenerational wealth allocation [87]. Given that these inequalities persist in both data collection and the data itself [88], potentially in implicit ways, any AMA reliant on its environment has the potential to propagate or even amplify these inequalities. As computer scientists, we have the opportunity to build algorithmic mechanisms into our AMAs to counteract and help remedy these types of bias [89]. If we take this pro-active approach to equality, it becomes important to

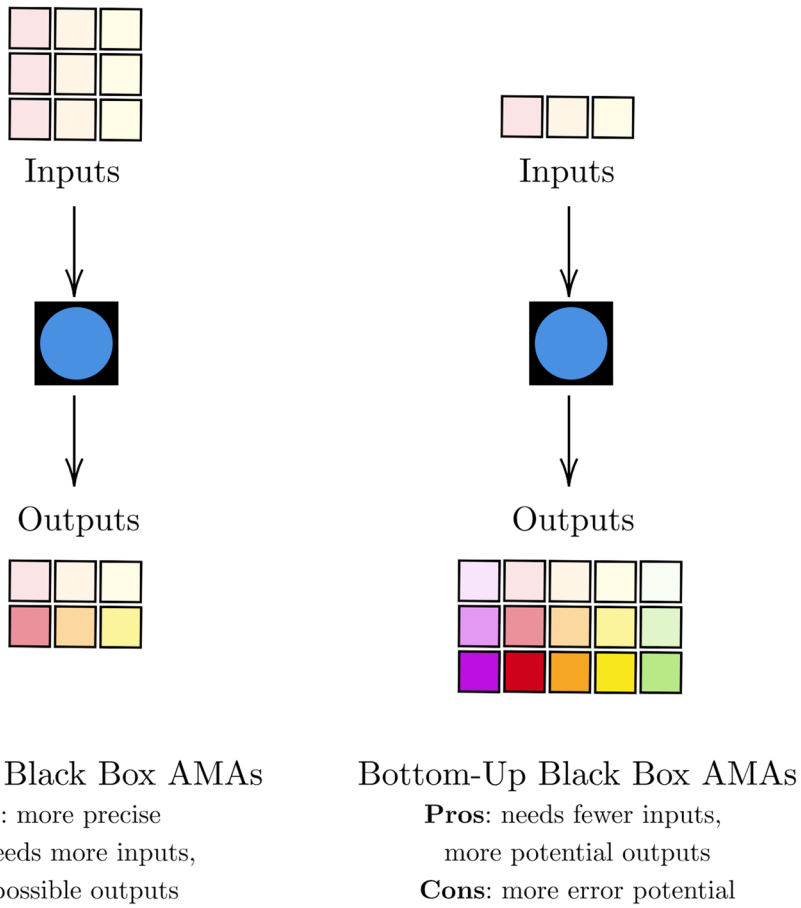


Fig. 5 Pros and cons of using black box AMAs. Black squares with a blue circle are black box AMAs (color figure online)

understand how biases exist in our environment, how these get encoded in our data, and how the AMA can use them inappropriately in its reasoning or explanation mechanisms [90]. This is more important in BU systems due to their higher capacity for learning morality from the environment [13].

For systemic inequalities that affect marginalised communities, minimising predictive disparities over different demographics is a proxy for our AMA "learning" appropriate moral principles. This line of work has been explored extensively in the fairness and sequential decision-making literature [12] and we briefly review important instances. Ensemble models have proven effective, with different weighting schemes used per classifier [83, 91] and for "unfairly classified" samples [92]. Coston and colleagues [93] found that characterising properties over the Rashomon set—an ensemble of models that all perform highly which, in this case, is the set of most fair models—gave them algorithmic bounds for the range of disparities with applications to recidivism risk prediction and consumer lending. Beyond outcome evaluation, Dai et al. [94] laid out metrics for evaluating post-hoc explanations: fidelity, stability, and consistency, aligning with the main conceptual challenges for IML [19]. They also evaluated the practical explanation quality of sparsity, which is a proxy for how "understandable" an explanation is, with higher sparsity allowing for fewer features and thus easier understanding. These ideas of human-intuitive explanations are expanded on in Sect. 4.2. Well-aligned explanations are useful because they can reduce overreliance on AI systems and make human-AI interaction more coordinated [95].

Taking a more engineering-based approach, Shaw et al. [96] defined meta-qualities for desired moral standards to guide BU AMAs under very constrained applications. Accepting that uncovering the reasoning of individual agents is challenging, the authors analysed multi-agent systems via post-hoc explainability to derive bounds for moral behaviour. These multi-agent behaviours can also be viewed from a TD lens via consideration of TD agents and stakeholders in a complete sociotechnical system [97] or collective actions of TD agent systems called moral communities [98]. The bounds in these cases are based on average population behaviour and act like probabilistic alternatives to the Moral Turing Test. Stronger fairness tests, such as those based on localised program execution paths [99], would be needed for individual instances of discrimination. A more intuitive way to view these "moral behaviours" is as highly likely probabilistic constraints—similar to the way we would view

the chance of a bridge retaining its shape under stress—they should hold under all reasonable perturbations within a given context.

The limitation of TD AMAs is that it is challenging to select the appropriate moral paradigm for a situation, made more difficult by the fact that there are several potentially appropriate moral decisions (i.e. input–output relation cardinality of moral decisions is one-to-many) based on varying sequential decision trajectories. Reinforcement learning (RL) has become the de facto toolkit for sequential decision-making since it can comprehensively explore a given decision space [12]—in moral agent terms, this amounts to BU flexibility within wide-ranging but well-defined TD constraints, or the hybrid moral paradigm. For robustness to this decision trajectory variation, Svegliato et al. [100] developed an AMA based on RL that uses all previous decision states to make its final decision. Notably, this AMA also circumvented the issue of imprecise objective functions by decoupling the moral compliance objectives from the task objectives. This decoupling was also recommended as one of the main ways to circumvent the Interpretation Problem in [72] and mirrors their distinction between moral mistakes and amoral mistakes. This avoids issues of ambiguous fidelity with respect to explanations [19]. In a less constrained decision-space than the setting in Svegliato et al. [100], trajectories can get more unwieldy. Subsequent work by Srivastava et al. [101] applies additional constraints to the agent (instead of its environment) which can analyse negative side effects of prototype decision sequences with human input and then replan an appropriate sequence which minimises these side effects. RL can also modify the environment to make it more appropriate. Using a reward that includes both the performance and moral compliance objectives, Rodriguez-Soto et al. [102] refine the convex hull of a Markov decision process (the stochastic process underlying RL) to get moral bounds on the decision space, which also mirrors the second suggestion for circumventing the Interpretation Problem.

Having begun this section with a discussion of fairness requirements, we covered how they can be enabled through metrics on the black box outputs, surveyed moral explanations for black boxes, and finished by reviewing high capacity black boxes for sequential decision-making. Regardless of the chosen interpretable black box paradigm, AMAs can be trusted without transparency if their power or scale is well-tuned to their purpose (as per Fig. 3). However, steps should be taken to enable debugging where applicable. Fairness and moral compliance objectives should be distinct from performance objectives [103, 104], and ideally, post-hoc explanations should be used to facilitate easier debugging in case of faulty agents. Even after deployment in real-world settings, AMAs and their explanations require ongoing evaluation and maintenance. Microsoft's Tay chatbot was an example of a BU agent that was initially safe but then exploited by social media users to consistently produce racist responses and conspiracy theories [105]. Similarly, Google's Gemini model caused controversy when a top-down goal of reductively prioritising diverse images led it to generate historically-inaccurate pictures [106]. As a general rule, explanations or decision trajectories should be stress-tested in different scenarios and their consistency analysed [19]. Thus, for additional safety, we recommend that the MLI for trustworthy black box models be consistent explanations or decision trajectories over important subgroups of the populations in the dataset. These model qualities are shown in Fig. 6.

4 Does transparency help us understand AMAs?

In this section, we frame interpretability as transparency for internal model reasoning, looking at two forms of transparency: **algorithmic transparency** and **decomposability** [18]. With that in mind, we show below that the utility of transparency varies in magnitude and is context-dependent. We propose algorithmic transparency as single-agent rules composed into multi-agent rules. Furthermore, we argue that the importance of transparency rises with the causal power of the agent, and depends on the relevant stakeholders.

In real-world settings, AMAs require higher capacity to interact and respond to their environment [72]. For this, we assume that the moral paradigm (BU/TD/hybrid) of the agent is sufficiently flexible to allow adaptation to new environments. With that assumption, the most important AMA construction parameters for analysing deployed AMAs become the scale and purpose of the agent. We note these are both somewhat nebulous terms that incorporate aspects outside AMA construction: 'scale' encompasses the number of model parameters, the capacity to act in the world, and the number of other agents which interact with it (for instance its users) and 'purpose' can be defined a priori by developers via its task objectives and moral paradigm but is ultimately at the hands of the user. Additionally, the purpose of an AMA is dynamic and can change with regards to performance capabilities achieved at scale. For added clarity when considering both construction parameters and transparency terms in the following sections, we centre the discussion of algorithmic transparency on AMA construction since specifics of the explanation form are less important, and we centre decomposability on AMA users since the explanation form is paramount to its utility.

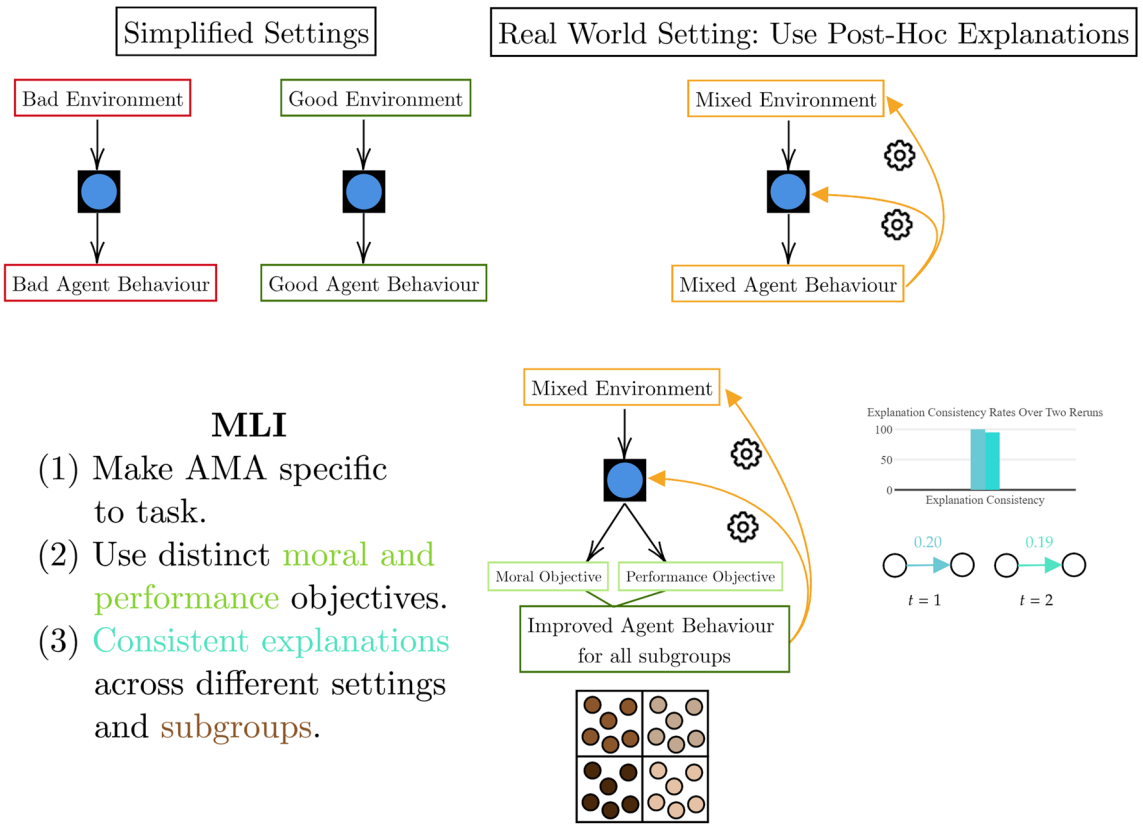


Fig. 6 Summary of black box AMAs and their MLI. Black squares with a blue circle are black box AMAs (color figure online)

4.1 The utility of algorithmic transparency depends on AMA construction

The Artus-Badea law states that an increase in scale gives an AMA more causal power and so more exposure to risk [72]. Thus, if an AMA does not have a significant effect on the world around it, there is less of a safety requirement for transparency to understand the agent's reasoning [63]. Therefore, in such trivial cases, one could get away with not implementing any explicit transparency. But moral "significance" is not always obvious because of collective agent behaviour from horizontally-scaled systems. For example, say you design an agent for your own use to analyse the sentiment of current news (whether the news is positive or negative) so you can prioritise more positive news stories. This only involves you and is thus unlikely to have a direct and significant impact on society, and would be a safe AMA without the requirement of transparency. However, if you decide to scale the model up or make it into a commercial product so that other people use the same agent, then moral questions arise because its effects on society are compounded and individuals might experience different uni-agent effects. For example, we have the development of echo chambers within social media websites, the reinforcement effect this has on sub-populations (perpetuating their existing opinions), and then the combined polarising effect on the entire population (the radicalisation of opposite sides) [107].

Now onto our proposed MLI for this case. As discussed in Sect. 3.2, multi-agent system behaviour used as post-hoc explainability can help give us guarantees on general agent morality *for specific tasks*. However, without some level of individual agent transparency, we have no guarantees on agent subgroup behaviour below a certain (unknown) subgroup size, and consequently a deeper understanding of the overall population becomes intractable. Internal ensembling over outputs is a practical way to get probabilistic limits to behaviour [108] for black box singular agents and to mitigate this population-subgroup mismatch. We discussed how Coston et al. [93] used ensembles over function space (i.e. potentially vastly different models) to aid in algorithmic fairness but Barnett et al. [109] propose a simpler alternative within a reinforcement learning framework that ensembles over two models predicting opposite things. They compare the output of rational and irrational teacher models to quantify the difference between them and thus produce a better "rationality direction" for future decision trajectories. Accordingly, we recommend the MLI for horizontally-scaled AMAs to be a nested combination of black box

interpretability. Thus our proposal is thinking of algorithmic transparency as logical rules [103, 110] (or as probabilistic limits [96, 103]) intended for uni-purpose individual agents and composed into multi-agent rules (or limits). In other words, this means bounding multi-agent systems by bounding uni-agent systems with algorithmic transparency. As the agents themselves scale vertically (i.e. their number of parameters increases), the range of behaviours each agent can perform increases and, when combined in multi-agent systems, results in complex collective behaviour [111]. To limit the complexity of our discussion, we do not talk about this combined vertical and horizontal scaling case further, instead moving to focusing on vertically-scaled agents.

Vertical scaling of models is performed through increasing three parameters: training data size (with the assumption that quality stays the same), compute power, and model capacity [112]. Although algorithmic transparency has not been studied for vertically scaled AMAs directly [13], we can use large language models (LLMs) as an approximation given their ongoing integration into MDM settings like medicine [113, 114]. LLMs such as GPT-3 have demonstrated linear scaling laws for prediction, that is: as the three parameters increase, the performance of the model increases linearly [112]. However, Ganguli et al. [115] showed that although these scaling laws hold at a general level across multiple tasks, performance on specific tasks can change abruptly at arbitrary scaling points of the three parameters, raising questions on what the models are actually learning. By reverse engineering neural networks, we are beginning to mathematically understand these flows of information [116] and discontinuous jumps to qualitatively better performance [117] during the learning process. However, these results are currently limited to toy models (neural networks with only one hidden layer or one attention module) of much smaller size than those in deployment and there is thus uncertainty about their generalisation to real-world models with multiple components [117]. Regardless, the rapid uptake and potential of these models necessitates guarantees on MDM to lessen harmful effects on end-users [118]. Where these guarantees are not currently expressible as a proof or formula, we propose expressing them as probabilistic guarantees or qualitative explanations that reveal opaque agent reasoning. These ideas are summarised in Fig. 7. As such, we continue the discussion of vertically-scaled agents via decomposability in large language models in the following section.

4.2 The utility of decomposability depends on the stakeholder

The main benefit of making interpretability accessible and intuitive to non-experts is that, when interpretability is viewed as a debugging tool for AMAs, it allows anyone to provide feedback for an AMA in order to improve its behaviour. Opening up feedback mechanisms at a societal level helps us create AMAs that are better-aligned with a diverse set of human values, and thus generalise better to unseen environments. The main challenge is that non-experts naturally require a simpler explanation than an expert. Not only is it difficult to get simple explanations for complex AMA reasoning, but there is a trade-off between simplicity and accuracy of an explanation [20]. Simpler explanations may be more understandable to non-expert users of AMAs but, if they become too simple, they can become inaccurate (as shown in Fig. 2) and the non-expert would not be aware of this. So, what makes an explanation of an agent's decision-making intuitive? As has been a common theme throughout this paper, there is no single answer since intuitive explanations are dependent on the type of stakeholders and their

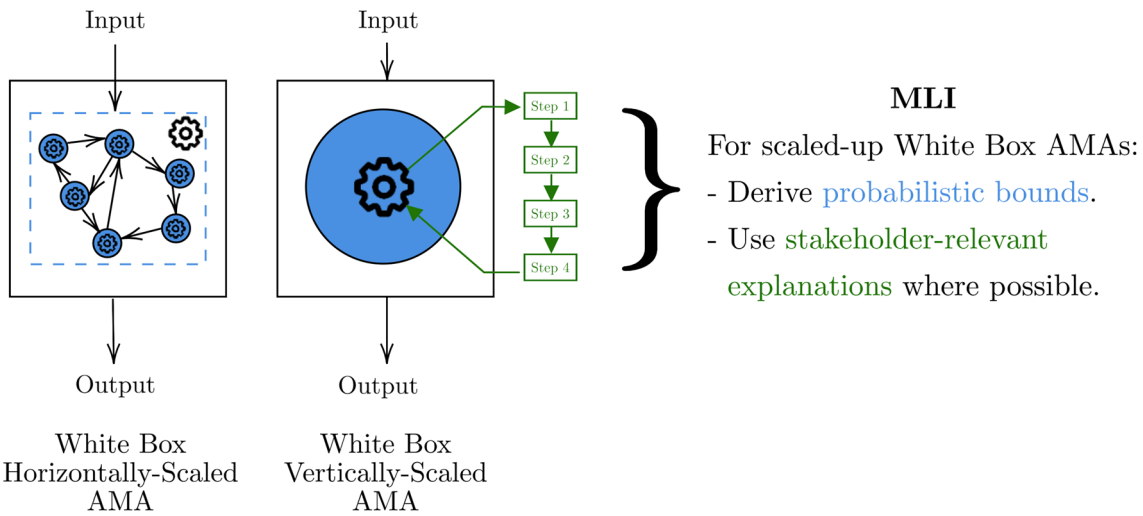


Fig. 7 Summary of AMAs discussed in Sect. 4 along with their MLI. White squares with a blue circle are white box AMAs (color figure online)

level of interaction with the agent. Suresh et al. [40] ascribe two essential facets to stakeholders in interpretability: their level of expertise (knowledge within a context), and their goals in the long- and short-term. While long-term goals (model understanding and trust) are common across all stakeholders, short-term goals can differ. Given our focus on MDM and user accessibility, we are focused on the (O1), (O7), and (T1–4) short-term goals (as defined in [40]) which revolve around model debugging, improvement, and feature importance [40, 119].

As part of pursuing these goals, we segment our stakeholders into two functional categories—developers and end users—for a clear and concrete analysis of stakeholder concerns in interpretable MDM. Intuitive explanations of agent reasoning steps are important for developers and users alike so that both can easily modify the AMA to rectify and prevent unintended behaviour (i.e. error reduction) while also fine-tuning the agent for more desirable behaviour (i.e. agent optimisation). To better align the actions of high-capacity AMAs to human values, we have touched on algorithmic mechanisms of counteracting bias and designing modular objective functions (moral and performance), but these interventions are restricted to developers. To make alignment accessible to the end users, we can instead express these algorithmic changes directly through text or other intuitive modality forms for humans (e.g. audio, video) [120]. One advantage is that this can help level the commercial and regulatory playing field between developers and users. Rudin [20] describes how the black box nature of models allows companies and their developers to get away with some faulty individual predictions if their average behaviour is sufficient, which the above proposal mitigates. We propose integrating decomposability into AMAs, which would mean that explanations would be presented in a way that allows users to understand the AMA's reasoning, giving them more control and customisation over the agent, and decreasing the chance that developers would be able to exploit them through asymmetric information [63].

In the same vein, we believe that the success of publicly-deployed LLMs like ChatGPT is largely due to their use of text as an interface for users. The input text for LLMs is called a "prompt"—the ease of prompt design and its interpretations as both probabilistic and textual inputs have inspired new ways of formalising LLMs at multiple interpretability scales [121, 122]. A key paper from this line of research is Chain-of-Thought (CoT) prompting [123] which involves setting up the prompt with sequential reasoning steps (a "rationale chain") to lead the LLM to give its response using a similar rationale chain. Extensions to CoT have focused on automating rationale chain generation via pre-defined prompt phrases to generate multiple rationale chains [124] and smart pruning of less likely rationale chains [125]. However, CoT reasoning is an emergent phenomenon of large models (>100 billion parameters). For medium-sized LLMs (10–100 billion parameters), the AI safety company Anthropic have found that they can learn moral concepts related to harm like stereotyping and bias when given clear instructions [126]. Going one step further, Jin et al. [47] developed a CoT extension which determines when it is appropriate to break moral rules with a provided rationale, giving textual explanations for both developers and users. Currently, this work is limited to three lower-stakes situations within the cultural context of the USA but the experiments are initial proof that medium-sized (and larger) LLMs are capable of learning and reasoning about moral obligations [79, 127, 128]. This suggests that larger LLMs can also be fine-tuned for specific moral consideration by all stakeholders with appropriate datasets.

The arguments above show once again the value of our proposal of using decomposability to enhance stakeholder accessibility and enable them to do more. Having explained the importance of decomposability, we also describe its fundamental limitation: *oversimplification*. Humans are complex adaptive systems that exist within the dynamics of societal interaction [39]. This complexity means digitisation (or conversion into "form" [74]) of context-dependent human concepts like trust, understanding, or morality is just an approximation of the real thing [39]. Explaining these digitised concepts, whether internally during processing or post-hoc, will only also be further approximations because explanations are simplified representations of the original model [20, 80]. Inputs to the agent need to contain causal information for outputs, otherwise explanations of morality cannot translate from humans to machines and back without a loss of information, making them brittle and particularly harmful to long-term trust [129]. This may seem like a pessimistic view of interpretability, but it draws attention to the fact that our internal representations of moral principles and their subsequent preservation through AMA processing are the keys to ensuring useful explanations. Referring back to our discussion of form and meaning in Sect. 3.1, useful model decomposition will require adaptive multi-modal agents grounded in experience of the world [75, 130].

5 Discussion

On the one hand, artificial moral agents (AMAs) can be created without interpretability and be given sufficient trust for moral reasoning in some narrow and well-defined tasks. The Interpretation Problem [72] means that we would still never get perfect guarantees about moral behaviour but we can get around this by building value-based agents that can be tested for trustworthiness [54]. However, relying on "input–output" tests, like the Moral Turing Test, limits this trust because they do not evaluate for "intrinsic" morality

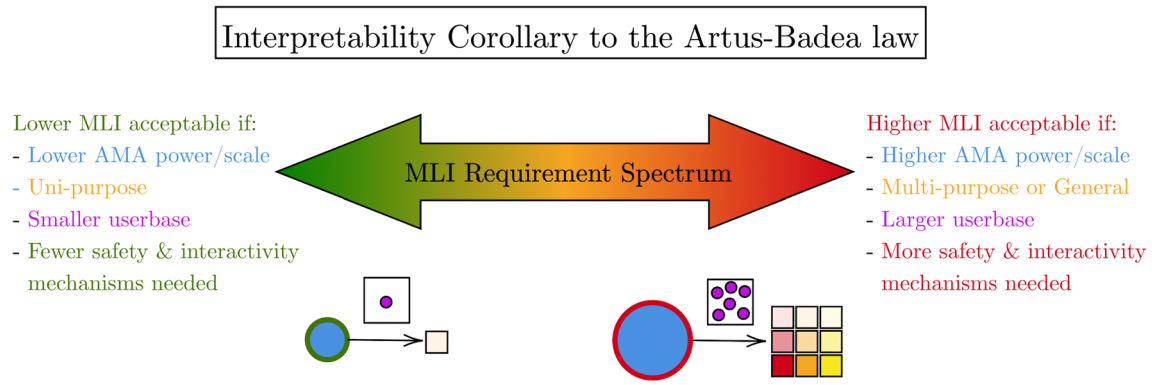


Fig. 8 Our proposed Interpretability Corollary to the Artus-Badea law [72] that summarises this paper

or the possibility of several acceptable moral outputs to the same inputs. To improve this and make individual agents more reliable, we can use principles based on collective agent behaviour to guide them and address this issue of trust in black box bottom-up agents with opaque internal mechanisms [96]. Additionally, top-down black box AMAs allow us to define prior moral constraints and carefully construct objective functions in AMAs so their moral reasoning is more consistent and predictable [100].

On the other hand, for general-purpose AMAs, we need stronger levels of interpretability requirements. While we have seen large-scale studies that show it is possible to reliably obtain good general AMA behaviour, a poor understanding of the inner mechanics of these agents has still resulted in abrupt scaling issues and unintended individual agent behaviour [115, 118]. For optimal levels of trust in and between agent-developer-user systems, explanations at different levels of abstraction, while imperfect [20, 39], are imperative to help these AMAs reach safe deployment. Importantly, for future work in this area, better quantification of moral compliance can decisively aid the understanding of interpretability requirements in different contexts, neuro-symbolic methods can help construct top-down and hybrid AMAs [131, 132], and causality across the three rungs of Pearl's ladder [133] will allow us to identify the key drivers of specific moral decisions [134–137].

In conclusion, while trustworthy AMAs can be created without any level of transparency, they make rigorous assessment and risk mitigation of AMAs much more challenging. For the moral paradigm of an AMA, we recommend top-down or hybrid agents, advocating against the use of bottom-up agents due to their higher risk of learning improper moral principles when deployed in novel environments. For AMAs with variable purposes, we believe algorithmic behavioural guarantees are the MLI for *uni-purpose AMAs*, with additional during-processing explanations, or task-specific decomposability, being the MLI for *general-purpose AMAs*. Additionally, for both of these purposes, moral compliance objectives should be as distinct from performance objectives as possible for easier "quantification" [19, 54]. For scale, both horizontally- and vertically-scaled systems require strong algorithmic behavioural guarantees, and for those with multiple stakeholders, intuitive explanations in both algorithmic and textual forms, or stakeholder-specific decomposability.

These arguments, as previously supported in the paper, are crystallised in our *Interpretability Corollary* to the Artus-Badea law [72]: the more power, or higher scale, an AMA has, and the more potential it has for a wide user base, the more safety and interactivity mechanisms are required, and thus a higher MLI is needed, as explored in Fig. 8. In other words: **with more power comes the requirement for a higher Minimum Level of Interpretability**.

Author contributions A.V. conceptualised the idea, did the literature review, and wrote the original draft. C.B. helped refine the structure of the paper and both authors contributed to subsequent drafts. Both authors have read and approved the final manuscript.

Funding A.V. is supported by a UK Research and Innovation (UKRI) Centre for Doctoral Training in AI for Healthcare PhD studentship (EPSRC, EP/S023283/1).

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Sallam, M.: Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare **11**(6), 887 (2023). https://doi.org/10.3390/healthcare11060887
2. Eloundou, T., Manning, S., Mishkin, P., Rock, D.: Gpts are gpts: an early look at the labor market impact potential of large language models (2023). arXiv arXiv:2303.10130
3. Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., Hu, X.: Harnessing the power of llms in practice: a survey on chatgpt and beyond (2023). arXiv arXiv:2304.13712
4. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-maron, G., Giménez, M., Sulsky, Y., Kay, J., Springenberg, J.T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., Freitas, N.: A generalist agent. Transactions on Machine Learning Research (2022). https://openreview.net/forum?id=1ikK0kHjvj
5. Ibarz, B., Kurin, V., Papamakarios, G., Nikiforou, K., Bennani, M., Csordás, R., Dudzik, A.J., Bošnjak, M., Vitvitskyi, A., Rubanova, Y.: A generalist neural algorithmic learner. In: Proceedings of the First Learning on Graphs Conference (2022). https://openreview.net/forum?id=FebadKZf6Gd
6. Jablonka, K.M., Schwaller, P., Smit, B.: Is gpt-3 all you need for machine learning for chemistry? In: NeurIPS 2022 Workshop on AI for Accelerated Materials Design (2022). https://openreview.net/forum?id=dgpgTEZ6G__
7. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022). https://aclanthology.org/2022.emnlp-main.256
8. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nat. Med. **28**(9), 1773–1784 (2022). https://doi.org/10.1038/s41591-022-01981-2
9. Mostafa, S.A., Ahmad, M.S., Mustapha, A.: Adjustable autonomy: a systematic literature review. Artif. Intell. Rev. **51**(2), 149–186 (2019). https://doi.org/10.1007/s10462-017-9560-8
10. Cervantes, J.-A., Rodríguez, L.-F., López, S., Ramos, F., Robles, F.: Autonomous agents and ethical decision-making. Cogn. Comput. **8**, 278–296 (2016). https://doi.org/10.1007/s12559-015-9362-8
11. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al.: Augmented language models: a survey (2023). arXiv arXiv:2302.07842
12. Nashed, S.B., Svegliato, J., Blodgett, S.L.: Fairness and sequential decision making: limits, lessons, and opportunities (2023). arXiv arXiv:2301.05753
13. Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: a survey of the current status. Sci. Eng. Ethics **26**, 501–532 (2020). https://doi.org/10.1007/s11948-019-00151-x
14. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017). https://doi.org/10.1089/big.2016.0047
15. Fatemi, M., Killian, T.W., Subramanian, J., Ghassemi, M.: Medical dead-ends and learning to identify high-risk states and treatments. In: Advances in Neural Information Processing Systems, vol. 34 (2021). https://proceedings.neurips.cc/paper_files/paper/2021/hash/26405399c51ad7b13b504e74eb7c696c-Abstract.html
16. Brutzman, D., Blais, C.L., Davis, D.T., McGhee, R.B.: Ethical mission definition and execution for maritime robots under human supervision. IEEE J. Ocean. Eng. **43**(2), 427–443 (2018). https://doi.org/10.1109/JOE.2017.2782959
17. Haidt, J.: Morality. Perspect. Psychol. Sci. **3**(1), 65–72 (2008). https://doi.org/10.1111/j.1745-6916.2008.00063.x
18. Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018). https://doi.org/10.1145/3236386.3241340
19. Watson, D.S.: Conceptual challenges for interpretable machine learning. Synthese **200**(2), 65 (2022). https://doi.org/10.1007/s11229-022-03485-5
20. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
21. Martinho, A., Poulsen, A., Kroesen, M., Chorus, C.: Perspectives about artificial moral agents. AI Ethics **1**(4), 477–490 (2021). https://doi.org/10.1007/s43681-021-00055-2
22. Hammond, L., Belle, V.: Learning tractable probabilistic models for moral responsibility and blame. Data Min. Knowl. Discov. **35**(2), 621–659 (2021). https://doi.org/10.1007/s10618-020-00726-4
23. Garrigan, B., Adlam, A.L., Langdon, P.E.: Moral decision-making and moral development: toward an integrative framework. Dev. Rev. **49**, 80–100 (2018). https://doi.org/10.1016/j.dr.2018.06.001
24. Piaget, J.: The Moral Judgement of the Child. Penguin (1932)
25. Kohlberg, L.: Moral stages and moralization: the cognitive-development approach. In: Moral Development and Behavior: Theory, Research and Social Issues, pp. 31–53 (1976)
26. Rest, J.R., Thoma, S.J., Bebeau, M.J., et al.: Postconventional Moral Thinking: A Neo-Kohlbergian Approach. Psychology Press (1999)
27. Gibbs, J.C.: Moral Development and Reality: Beyond the Theories of Kohlberg, Hoffman, and Haidt. Oxford University Press (2013). https://doi.org/10.1093/acprof:osobl/9780199976171.001.0001
28. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. **108**(4), 814 (2001). https://doi.org/10.1037/0033-295X.108.4.814
29. Haidt, J., Bjorklund, F.: Social intuitionists answer six questions about morality. Moral Psychology (2008). https://ssrn.com/abstract=855164
30. Kahneman, D.: Thinking, Fast and Slow. Macmillan (2011)
31. Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D.: An fmri investigation of emotional engagement in moral judgment. Science **293**(5537), 2105–2108 (2001). https://doi.org/10.1126/science.1062872
32. Greene, J., Haidt, J.: How (and where) does moral judgment work? Trends Cogn. Sci. **6**(12), 517–523 (2002). https://doi.org/10.1016/S1364-6613(02)02011-9
33. Gauthier, D.: Morals by Agreement. Clarendon Press (1987). https://doi.org/10.1093/0198249926.001.0001
34. Vitell, S.J., Nwachukwu, S.L., Barnes, J.H.: The effects of culture on ethical decision-making: an application of Hofstede's typology. J. Bus. Ethics **12**, 753–760 (1993). https://doi.org/10.1007/BF00881307
35. Cribb, A., Entwistle, V.A.: Shared decision making: trade-offs between narrower and broader conceptions. Health Expect. **14**(2), 210–219 (2011). https://doi.org/10.1111/j.1369-7625.2011.00694.x
36. Berman, J.Z., Kupor, D.: Moral choice when harming is unavoidable. Psychol. Sci. **31**(10), 1294–1301 (2020). https://doi.org/10.1177/0956797620948821
37. Helberger, N., Araujo, T., Vreese, C.H.: Who is the fairest of them all? public attitudes and expectations regarding automated

decision-making. Comput. Law Secur. Rev. **39**, 105456 (2020). https://doi.org/10.1016/j.clsr.2020.105456
38. Asch, S.E.: Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychol. Monogr. Gen. Appl. **70**(9), 1 (1956). https://doi.org/10.1037/h0093718
39. Birhane, A.: The impossibility of automating ambiguity. Artif. Life **27**(1), 44–61 (2021). https://doi.org/10.1162/artl_a_00336
40. Suresh, H., Gomez, S.R., Nam, K.K., Satyanarayan, A.: Beyond expertise and roles: a framework to characterize the stakeholders of interpretable machine learning and their needs. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (2021). https://doi.org/10.1145/3411764.3445088
41. Cai, C.J., Winter, S., Steiner, D., Wilcox, L., Terry, M.: "hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making (2019). https://doi.org/10.1145/3359206
42. Feng, L., Wiltsche, C., Humphrey, L., Topcu, U.: Synthesis of human-in-the-loop control protocols for autonomous systems. IEEE Trans. Autom. Sci. Eng. **13**(2), 450–462 (2016). https://doi.org/10.1109/TASE.2016.2530623
43. Araujo, T., Helberger, N., Kruikemeier, S., De Vreese, C.H.: In ai we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc. **35**, 611–623 (2020). https://doi.org/10.1007/s00146-019-00931-w
44. Upton, C.L.: Virtue ethics and moral psychology: the situationism debate. J. Ethics **13**(2–3), 103–115 (2009)
45. Hindocha, S., Badea, C.: Moral exemplars for the virtuous machine: the clinician's role in ethical artificial intelligence for healthcare. AI Ethics **2**(1), 167–175 (2022). https://doi.org/10.1007/s43681-021-00089-6
46. Post, B., Badea, C., Faisal, A., Brett, S.J.: Breaking bad news in the era of artificial intelligence and algorithmic medicine: an exploration of disclosure and its ethical justification using the hedonic calculus. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00230-z
47. Jin, Z., Levine, S., Gonzalez Adauto, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., Schölkopf, B.: When to make exceptions: exploring language models as accounts of human moral judgment. In: Advances in Neural Information Processing Systems, vol. 35 (2022). https://openreview.net/forum?id=uP9RiC4uVcR
48. Mattingly, C., Throop, J.: The anthropology of ethics and morality. Annu. Rev. Anthropol. **47**, 475–492 (2018). https://doi.org/10.1146/annurev-anthro-102317-050129
49. Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. **21**(4), 18–21 (2006). https://doi.org/10.1109/MIS.2006.80
50. Allen, C., Smit, I., Wallach, W.: Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics Inf. Technol. **7**, 149–155 (2005). https://doi.org/10.1007/s10676-006-0004-4
51. Provost, F.J., Hennessy, D.N.: Scaling up: Distributed machine learning with cooperation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 13 (1996)
52. Ali, A.H.: A survey on vertical and horizontal scaling platforms for big data analytics. Int. J. Integr. Eng. **11**(6), 138–150 (2019). https://doi.org/10.30880/ijie.2019.11.06.015
53. Parthemore, J., Whitby, B.: What makes any agent a moral agent? Reflections on machine consciousness and moral agency. In. J. Mach. Conscious. **5**(02), 105–129 (2013)
54. Badea, C.: Have a break from making decisions, have a mars: the multi-valued action reasoning system. In: Artificial Intelligence XXXIX: 42nd SGAI International Conference on Artificial Intelligence (2022). https://doi.org/10.1007/978-3-031-21441-7_31
55. Floridi, L., Sanders, J.W.: On the morality of artificial agents. Minds Mach. **14**, 349–379 (2004)
56. Johnson, D.G.: Computer systems: moral entities but not moral agents. Ethics Inf. Technol. **8**, 195–204 (2006)
57. Brożek, B., Janik, B.: Can artificial intelligences be moral agents? New Ideas Psychol. **54**, 101–106 (2019)
58. Behdadi, D., Munthe, C.: A normative approach to artificial moral agency. Minds Mach. **30**(2), 195–218 (2020)
59. Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D.P., Wilson, A.G.: Loss surfaces, mode connectivity, and fast ensembling of dnns. In: Advances in Neural Information Processing Systems, vol. 31 (2018). https://papers.nips.cc/paper_files/paper/2018/hash/be3087e74e9100d4bc4c6268cdbe8456-Abstract.html
60. Fort, S., Dziugaite, G.K., Paul, M., Kharaghani, S., Roy, D.M., Ganguli, S.: Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In: Advances in Neural Information Processing Systems, vol. 33 (2020). https://proceedings.neurips.cc/paper/2020/hash/405075699f065e43581f27d67bb68478-Abstract.html
61. Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: convergence and generalization in neural networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018). https://papers.nips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html
62. Krishnan, M.: Against interpretability: a critical examination of the interpretability problem in machine learning. Philos. Technol. **33**(3), 487–502 (2020). https://doi.org/10.1007/s13347-019-00372-9
63. Molnar, C.: Interpretable machine learning. Lulu.com (2020). https://christophm.github.io/interpretable-ml-book/
64. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. **144**(1), 114 (2015). https://doi.org/10.1037/xge0000033
65. Johansson, L.: The functional morality of robots. Int. J. Technoethics **1**(4), 65–73 (2010). https://doi.org/10.4018/jte.20101001 05
66. Allen, C., Varner, G., Zinser, J.: Prolegomena to any future artificial moral agent. J. Exp. Theor. Artif. Intell. **12**(3), 251–261 (2000). https://doi.org/10.1080/09528130050111428
67. Searle, J.R.: Minds, brains, and programs. Behav. Brain Sci. **3**(3), 417–424 (1980). https://doi.org/10.1017/S0140525X00005756
68. Skorupski, J.: The definition of morality. R. Inst. Philos. Suppl. **35**, 121–144 (1993). https://doi.org/10.1017/S1358246100006299
69. Hardy, S.A., Carlo, G.: Moral identity: what is it, how does it develop, and is it linked to moral action? Child. Dev. Perspect. **5**(3), 212–218 (2011). https://doi.org/10.1111/j.1750-8606.2011.00189.x
70. Prior, W.J.: Can virtue be taught? Laetaberis J. Calif. Class. Assoc. **8**(1), 1–16 (1990-91)
71. Straughan, R.: Can we Teach Children to be Good? Basic Issues in Moral, Personal, and Social Education. McGraw-Hill Education (1988)
72. Badea, C., Artus, G.: Morality, machines, and the interpretation problem: a value-based, Wittgensteinian approach to building moral agents. In: Artificial Intelligence XXXIX: 42nd SGAI International Conference on Artificial Intelligence (2022). https://doi.org/10.1007/978-3-031-21441-7_9
73. Hoque, E.: Memorization: a proven method of learning. Int. J. Appl. Res. **22**(3), 142–150 (2018)
74. Bender, E.M., Koller, A.: Climbing towards nlu: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5185–5198 (2020). https://doi.org/10.18653/v1/2020.acl-main.463

75. Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A.: Experience grounds language. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8718–8735 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.703
76. Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P.: Towards out-of-distribution generalization: a survey (2021). arXiv/CoRR arXiv:2108.13624
77. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M.: Machine learning: the high interest credit card of technical debt. In: NeurIPS 2014 Workshop on Software Engineering for Machine Learning (SE4ML) (2014). https://papers.nips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf
78. Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Génova, G., Haim, G., Kließ, M.S., Lopez-Sanchez, M.: Ethics by design: necessity or curse? In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 60–66 (2018). https://doi.org/10.1145/3278721.3278745
79. Hendrycks, D., Burns, C., Basart, S., Critch, A.C., Li, J.L., Song, D., Steinhardt, J.: Aligning ai with shared human values. In: International Conference on Learning Representations, vol. 9 (2021). https://openreview.net/forum?id=dNy_RKzJacY
80. Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N.A., Choi, Y.: Social bias frames: reasoning about social and power implications of language. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5477–5490 (2020). https://doi.org/10.18653/v1/2020.acl-main.486
81. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., Floridi, L.: Ethics as a service: a pragmatic operationalisation of ai ethics. Minds Mach. **31**(2), 239–256 (2021). https://doi.org/10.1007/s11023-021-09563-w
82. Tilly, C.: Historical perspectives on inequality. In: The Blackwell Companion to Social Inequalities, pp. 15–30 (2005). https://doi.org/10.1002/9780470996973.ch2
83. Kenfack, P.J., Khan, A.M., Kazmi, S.A., Hussain, R., Oracevic, A., Khattak, A.M.: Impact of model ensemble on the fairness of classifiers in machine learning. In: 2021 International Conference on Applied Artificial Intelligence (ICAPAI), pp. 1–6 (2021). https://doi.org/10.1109/ICAPAI49758.2021.9462068
84. Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A.R., Kuchenbaecker, K.: A roadmap to increase diversity in genomic studies. Nat. Med. **28**(2), 243–250 (2022). https://doi.org/10.1038/s41591-021-01672-4
85. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z
86. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax234
87. Heidari, H., Kleinberg, J.: Allocating opportunities in a dynamic model of intergenerational mobility. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 15–25 (2021). https://doi.org/10.1145/3442188.3445867
88. Ricaurte, P.: Data epistemologies, the coloniality of power, and resistance. Telev. New Media **20**(4), 350–365 (2019)
89. Mohamed, S., Png, M.-T., Isaac, W.: Decolonial ai: decolonial theory as sociotechnical foresight in artificial intelligence. Philos. Technol. **33**(4), 659–684 (2020). https://doi.org/10.1007/s13347-020-00405-8
90. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P.: Towards a standard for identifying and managing bias in artificial intelligence. NIST Special Publication 1270, pp. 1–77 (2022). https://doi.org/10.6028/NIST.SP.1270
91. Gohar, U., Biswas, S., Rajan, H.: Towards understanding fairness and its composition in ensemble machine learning (2022). arXiv arXiv:2212.04593
92. Bhaskaruni, D., Hu, H., Lan, C.: Improving prediction fairness via model ensemble. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1810–1814 (2019). https://doi.org/10.1109/ICTAI.2019.00273
93. Coston, A., Rambachan, A., Chouldechova, A.: Characterizing fairness over the set of good models under selective labels. In: International Conference on Machine Learning, vol. 38, pp. 2144–2155 (2021). http://proceedings.mlr.press/v139/coston21a/coston21a.pdf
94. Dai, J., Upadhyay, S., Aivodji, U., Bach, S.H., Lakkaraju, H.: Fairness via explanation quality: evaluating disparities in the quality of post hoc explanations. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 203–214 (2022). https://doi.org/10.1145/3514094.3534159
95. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S., Krishna, R.: Explanations can reduce overreliance on ai systems during decision-making. Proc. ACM Hum.-Comput. Interact. **7**(CSCW1), 1–38 (2023). https://doi.org/10.1145/3579605
96. Shaw, N.P., Stöckel, A., Orr, R.W., Lidbetter, T.F., Cohen, R.: Towards provably moral ai agents in bottom-up learning frameworks. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 271–277 (2018). https://doi.org/10.1145/3278721.3278728
97. Murukannaiah, P.K., Ajmeri, N., Jonker, C.M., Singh, M.P.: New foundations of ethical multiagent systems. In: Proceedings of the 19th Conference on Autonomous Agents and MultiAgent Systems (2020)
98. Nashed, S., Svegliato, J., Zilberstein, S.: Ethically compliant planning within moral communities. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 188–198 (2021). https://doi.org/10.1145/3461702.3462522
99. Aggarwal, A., Lohia, P., Nagar, S., Dey, K., Saha, D.: Black box fairness testing of machine learning models. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 625–635 (2019). https://doi.org/10.1145/3338906.3338937
100. Svegliato, J., Nashed, S.B., Zilberstein, S.: Ethically compliant sequential decision making. In: AAAI Conference on Artificial Intelligence (AAAI) (2021). https://ojs.aaai.org/index.php/AAAI/article/view/17386
101. Srivastava, A., Saisubramanian, S., Paruchuri, P., Kumar, A., Zilberstein, S.: Planning and learning for non-Markovian negative side effects using finite state controllers. In: AAAI Conference on Artificial Intelligence (AAAI) (2023). https://ojs.aaai.org/index.php/AAAI/article/view/26767
102. Rodriguez-Soto, M., Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A.: Instilling moral value alignment by means of multi-objective reinforcement learning. Ethics Inf. Technol. **24**(1), 9 (2022). https://doi.org/10.1007/s10676-022-09635-0
103. Rossi, F., Mattei, N.: Building ethically bounded ai. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9785–9789 (2019). https://doi.org/10.1609/aaai.v33i01.33019785
104. Svegliato, J., Nashed, S., Zilberstein, S.: An integrated approach to moral autonomous systems. In: European Conference on Artificial Intelligence (ECAI) 325, pp. 2941–2942 (2020). https://doi.org/10.3233/FAIA200464

105. Learning from Tay's introduction. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/. Accessed 17 July 2024
106. Google to fix AI picture bot after 'woke' criticism. https://www.bbc.co.uk/news/business-68364690. Accessed 17 July 2024
107. Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proc. Natl. Acad. Sci. **118**(9), e2023301118 (2021). https://doi.org/10.1073/pnas.2023301118
108. Tekin, C., Yoon, J., Van Der Schaar, M.: Adaptive ensemble learning with confidence bounds. IEEE Trans. Signal Process. **65**(4), 888–903 (2016). https://doi.org/10.1109/TSP.2016.2626250
109. Barnett, P., Freedman, R., Svegliato, J., Russell, S.: Active reward learning from multiple teachers. In: AAAI 2023 Workshop on Artificial Intelligence Safety (SafeAI) (2023). arXiv:2303.00894
110. Mermet, B., Simon, G.: Formal verication of ethical properties in multiagent systems. In: 1st Workshop on Ethics in the Design of Intelligent Agents (2016). https://hal.science/hal-01708133/document
111. Kelly, K.: Out of Control: The New Biology of Machines, Social Systems, and the Economic World. Hachette UK (2009)
112. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020). arXiv arXiv:2001.08361
113. Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., Faix, D.J., Goodman, A.M., Longhurst, C.A., Hogarth, M., Smith, D.M.: Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern. Med. **183**(6), 589–596 (2023). https://doi.org/10.1001/jamainternmed.2023.1838
114. Lee, P., Bubeck, S., Petro, J.: Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. N. Engl. J. Med. **388**(13), 1233–1239 (2023). https://doi.org/10.1056/NEJMsr2214184
115. Ganguli, D., Hernandez, D., Lovitt, L., DasSarma, N., Henighan, T., Jones, A., Joseph, N., Kernion, J., Mann, B., Askell, A., et al.: Predictability and surprise in large generative models (2022). arXiv arXiv:2202.07785
116. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021). https://transformer-circuits.pub/2021/framework/index.html
117. Nanda, N., Chan, L., Lieberum, T., Smith, J., Steinhardt, J.: Progress measures for grokking via mechanistic interpretability. In: International Conference on Learning Representations, vol. 11 (2023). https://openreview.net/forum?id=9XFSbDPmdW
118. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al.: Ethical and social risks of harm from language models (2021). arXiv arXiv:2112.04359
119. Miller, G.J.: Stakeholder-accountability model for artificial intelligence projects. J. Econ. Manag. **44**(1), 446–494 (2022). https://doi.org/10.22367/jem.2022.44.18
120. Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., Wang, T.: A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. Decis. Support Syst. **152**, 113647 (2022). https://doi.org/10.1016/j.dss.2021.113647
121. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023). https://doi.org/10.1145/3560815
122. Dohan, D., Xu, W., Lewkowycz, A., Austin, J., Bieber, D., Lopes, R.G., Wu, Y., Michalewski, H., Saurous, R.A., Sohl-Dickstein, J.: Language model cascades (2022). arXiv:2207.10342
123. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems, vol. 35 (2022). https://openreview.net/pdf?id=_VjQlMeSB_J
124. Huang, J., Gu, S.S., Hou, L., Wu, Y., Wang, X., Yu, H., Han, J.: Large language models can self-improve (2022). arXiv arXiv:2210.11610
125. Shum, K., Diao, S., Zhang, T.: Automatic prompt augmentation and selection with chain-of-thought from labeled data (2023). arXiv arXiv:2302.12822
126. Ganguli, D., Askell, A., Schiefer, N., Liao, T., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., et al.: The capacity for moral self-correction in large language models (2023). arXiv arXiv:2302.07459
127. Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., Choi, Y.: Delphi: Towards machine ethics and norms (2021). arXiv arXiv:2110.07574
128. Forbes, M., Hwang, J.D., Shwartz, V., Sap, M., Choi, Y.: Social chemistry 101: learning to reason about social and moral norms. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 653–670 (2020). https://aclanthology.org/2020.emnlp-main.48/
129. Papagni, G., Köszegi, S.: Interpretable artificial agents and trust: supporting a non-expert users perspective. In: Culturally Sustainable Social Robotics, vol. 335 (2020). https://doi.org/10.3233/FAIA200974
130. Chandu, K.R., Bisk, Y., Black, A.W.: Grounding 'grounding' in nlp. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4283–4305 (2021). https://aclanthology.org/2021.findings-acl.375/
131. Sousa Ribeiro, M., Leite, J.: Aligning artificial neural networks and ontologies towards explainable ai. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 4932–4940 (2021). https://ojs.aaai.org/index.php/AAAI/article/view/16626
132. Roy, K., Gaur, M., Rawte, V., Kalyan, A., Sheth, A.: Proknow: process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance. Front. Big Data **5**, 1056728 (2022). https://doi.org/10.3389/fdata.2022.1056728
133. Pearl, J.: Causality. Cambridge University Press (2009). https://doi.org/10.1017/CBO9780511803161
134. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness, 30 (2017). https://papers.nips.cc/paper_files/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
135. Mhasawade, V., Chunara, R.: Causal multi-level fairness. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 784–794 (2021). https://doi.org/10.1145/3461702.3462587
136. Nashed, S.B., Mahmud, S., Goldman, C.V., Zilberstein, S.: A unifying framework for causal explanation of sequential decision making (2022). arXiv arXiv:2205.15462
137. Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., Potts, C.: Inducing causal structure for interpretable neural networks. In: International Conference on Machine Learning, vol. 39, pp. 7324–7338 (2022). https://proceedings.mlr.press/v162/geiger22a/geiger22a.pdf

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.