Original article

EXPERIMENTAL PHILOSOPHY

Is an Intelligent Machine a Moral Machine?

Simon Myers[1*], Jim A.C. Everett[2**]

[1] *Warwick Business School, Coventry, United Kingdom*
[2] *University of Kent, Canterbury, United Kingdom*

[*]Simon.Myers@wbs.ac.uk [**]j.a.c.everett@kent.ac.uk

Abstract

As artificial intelligence (AI) systems become increasingly sophisticated and used in more consequential domains, an unspoken assumption seems to suggest enhanced performance or "intelligence" would entail greater alignment and safety – "morality". And yet, as argued in the *Orthogonality Thesis* from AI ethics, an artificial agent becoming more intelligent does not mean it would necessarily become more moral, and therefore increased intelligence alone cannot reduce the danger that AI systems could pose. In this paper we draw on these philosophical debates and explore the psychological foundations of this apparent misconception: do people infer machine morality from machine intelligence? Across nine pre-registered studies (total n = 3895) we investigated how perceptions of AI intelligence shape perceived morality, and how increased AI intelligence shapes both perceptions of trustworthiness and safety in both human and artificial agents. While most work focuses on intelligence and morality as independent and orthogonal facets of person perception and trust, we instead highlight a robust pattern of results in which people do not only perceive intelligence and morality in AI agents, but concerningly infer moral competence and moral motivation from machine intelligence - with consequences for trust and danger. These findings reveal a systematic tendency to infer moral qualities from intelligence, including moral motivation. This may distort public understanding of AI safety and trust, with concerning ethical and epistemic implications.

Keywords Artificial Intelligence, Morality, Person Perception, Orthogonality Thesis, AI Safety

1. Introduction

Open almost any newspaper, article, or podcast and you will hear numerous claims about the rise in artificial intelligence and how it is reshaping the world around us. AI systems appear to be rapidly becoming more sophisticated and capable, sometimes far exceeding the expectations of their designers (Eschenbach, 2021). AI agents have been predicted to proliferate across many aspects of modern life and across many different domains, from finance to transport to crime prediction, and even to policymaking (Aharoni et al., 2022; Bonnefon et al., 2016; Cao, 2022; Esmaeilzadeh, 2020). AI optimists have welcomed these predictions, with some even suggesting

DOI: 10.18716/ojs/xphi/2026.1.11955

that before long the intelligence of these artificial agents may even *exceed* our own (Kannegieter, 2024; OpenAI, 2023; Roser, 2023). In contrast to such optimistic narratives, of course, others have argued more critically, that suggestions of a new AI revolution are overblown and misrepresents both what they can do now and what we should expect them to be able to do in the near future, (Dentella et al., 2024; Marcus, 2018, 2024; Narayanan & Kapoor, 2024).

While the AI optimists and the AI pessimists may disagree on the current state and trajectory of AI "intelligence", what there does seem to be agreement on is the importance of AI *safety* (Bostrom, 2012, 2014; Christian, 2021; Narayanan & Kapoor, 2024; Yudkowsky, 2016), and particularly how we can ensure that AI agents have values aligned with ours (Bostrom, 2014; Christian, 2021; Gabriel, 2020; Turchin, 2019; Yudkowsky, 2016). In order for an AI agent to be safe and worthy of our trust it has, to some extent, to care about the things we care about – to align with human moral values, have the same kind of concerns that we do, and be motivated to act in a moral way. This, of course, is necessary for AI to be trustworthy, rather than just being trusted. But as artificial agents increase in their "intelligence", do people expect them to increase in morality, and *should* they?

1.1 Two Dimensions in Social Cognition

In our social world, people need to form fast, global social impressions from limited information. In particular, they need to know the answers to two key questions: "Does this person wish me harm?", and "Can this person enact those intentions?" Decades of social psychology research has shown that we perceive others along two broad dimensions corresponding to these questions: one relating to morality or warmth, connecting sociality, good intentions, and moral character, the other relating to competence, the ability to achieve one's goals, including traits such as intelligence, skill, and talent (Abele & Wojciszke, 2007, 2014; Fiske et al., 2007; Wojciszke, 1994, 2005). More recent work has explicitly distinguished warmth from morality within the first dimension, showing that morality (traits like trustworthiness, honesty, and kindness) is particularly important for impression formation (Brambilla et al., 2021; Goodwin et al., 2014; Wojciszke et al., 1998). Similarly, models of trust distinguish between trust based on perceived competence or reliability and trust based on perceived integrity and benevolence (Barber, 1983; Deutsch, 1960; Mayer et al., 1995), sharing the idea that trust depends both on whether someone wishes us harm and whether they can enact it.

These two dimensions are typically treated as separable (e.g., Dubois & Beauvois, 2005; Fiske et al., 2007): someone can be kind but incompetent, or highly capable but selfish. And yet, there is evidence that perceptions of one dimension influence the other. Warmth and competence ratings can be positively correlated (Rosenberg et al., 1968), consistent with a halo effect (Nisbett & Wilson, 1977; Thorndike, 1920). More importantly, learning about moral character has a causal effect on attributions of competence. Stellar & Willer (2018) showed that people judged those who committed immoral acts as less competent, especially social competence, even while reporting that moral character should be irrelevant to job performance. Perceived social intelligence may act as a bridge between general competence and morality (Kim et al., 2004; Stellar & Willer, 2018), suggesting that some aspects of perceived morality rely on specific aspects of competence (e.g. one's ability to understand norms).

1.2 Intelligence and Morality in Artificial Agents

While decades of social cognition research have looked at perceptions of these two dimensions - henceforth intelligence and morality - as humans increasingly act with agents powered by artificial intelligence, we also need to understand how people perceive artificial agents and what the consequences of these intuitions are for trust and danger.

In line with the "Computers as Social Actors framework" (e.g., Nass et al., 1994; Reeves & Nass, 1996) there is evidence that people apply social rules and expectations to machines (Nass & Moon, 2000), perceive computer personalities (Nass & Moon, 2000), and appear to anthropomorphise AI in much the same way that they anthropomorphise other non-human agents (Epley et al., 2007). In particular, there is evidence that people form impressions of warmth ("morality") and competence ("intelligence") in artificial agents (McKee et al., 2023), and that these impressions of machine intelligence occur at the both the implicit and explicit level (Surdel et al., 2024).

In the context of trust, there has been increasing attention to how people's trust in AI agents is not driven only by perceptions of what the system can do, but why and how it does it (see Everett et al., 2026 for a recent review). When evaluating AI systems, trust is determined not only by perceptions of performance, or intelligence ("Does it work?") but also by moral considerations about the system's ethicality and sincerity ("Is it good?"). This multidimensionality of trust means that trust in AI depends not only on performance, ability, or intelligence, but also moral concerns about ethicality and sincerity (Everett et al., 2026; Lee & See, 2004; Malle & Ullman, 2021). For example, recent work looking at reported trust in both AI in general and 20 different specific systems shows that perceptions of both performance (reliability, competence) and morality (ethicality, sincerity) predicted overall trust in AI, both in general and for the specific systems (Claessens & Everett, 2026).

Related work has examined how the perceived competence of robots and AI systems affects moral evaluations of their *decisions*, showing that competence manipulations can moderate how moral an agent's actions are judged (Laakasuo et al., 2023, 2025), and that physical appearance rather than cognitive ability alone can drive moral condemnation (Malle et al., 2015; Sundvall et al., 2023). However, these studies examine moral evaluation of specific actions rather than the attribution of moral *traits*, that is, whether people come to see a more intelligent agent as a more moral *entity*, with greater moral competence and moral motivation.

While there is evidence that people infer cues about *both* AI intelligence and morality, and that these influence trust, it is much less clear whether people infer AI intelligence *from* morality, - or what the consequences for trust are. Researchers have explored attributions of intelligence from machine behaviour (De Neys & Raoelison, 2025), and AI research is forcing more precise definitions of what constitutes morality (Dahl, 2023; Purcell & Bonnefon, 2023), and how AI systems themselves align with human moral intuitions (Bonnefon et al., 2016; Zaim bin Ahmad & Takemoto, 2025). Yet we do not know how people infer intelligence from morality or vice versa. While there is relevant work from the enhancement literature, as with work in classic social cognition in humans, it remains unclear how these processes would occur with deliberately artificial agents. For example, Koverola et al. (2022) found that superhuman levels of cognitive ability achieved through neurotechnological enhancement were associated with perceived immorality and dehumanization, suggesting that moral evaluations of intelligence may depend on whether the intelligence is perceived as natural or artificially enhanced (see also Grinschgl et al., 2022, 2023). Given that AI systems are artificial by definition, this raises the intriguing possibility that people may infer morality from intelligence when thinking about artificial agents.

So do people infer morality from AI intelligence? This question is not only theoretically important for models of social cognition in the new domain of non-human agents, but also practically urgent for discussions about AI safety. Even if these two dimensions are broadly dissociable, we might still think there is some degree of a causal relationship between intelligence and morality in humans given our evolutionary history. The same, however, does not seem to be true for artificial agents. Most importantly, AI safety critics have proposed the opposite, that when it comes to AI intelligence and morality are *orthogonal*: a smarter machine need not be a safer, more ethical one (Armstrong, 2013; Bostrom, 2012). If laypeople infer morality from intelligence *per se*, rather than from the understanding that extra safety work is

necessary for moral alignment (and indeed becomes ever more necessary as models become more powerful), their trust in AI may rest on epistemically fragile foundations.

1.3 The Orthogonality Thesis

According to the "*Orthogonality Thesis*" (Armstrong, 2013; Bostrom, 2012), increased intelligence, capability, and likelihood of goal-achievement in a machine does not necessarily lead to greater morality or ethical sensitivity. A highly intelligent but unaligned AI, regardless of whether its goals are trivial, morally questionable, or even benevolent, could pose an existential threat to humanity. One of the reasons for this concern arises from the principle of *Instrumental Convergence* (Armstrong, 2013; Omohundro, 2018) which suggests that a wide variety of final goals will lead rational agents to adopt similar intermediate strategies, because certain instrumental goals (e.g., acquiring resources, seeking power, avoiding shutdown) are generally useful for achieving almost any terminal objective. Pursuing these goals without adequate constraint is likely to conflict with human values. As a result, an unaligned AI may not safeguard the things we care about, regardless of its level of intelligence.

The Orthogonality Thesis is not a claim about one necessarily observing statistical independence in reality, and it is fully compatible with the existence of a positive correlation between intelligence and moral alignment in practice (Bostrom, 2014). Indeed, there is growing evidence that in currently deployed AI systems, more capable models tend to be more aligned with human moral intuitions (Takemoto, 2026; Zaim Bin Ahmad & Takemoto, 2025; see also Anthropic, 2025). This observed correlation, however, is largely the product of deliberate and resource-intensive safety engineering, including reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ouyang et al., 2022), constitutional AI (Bai et al., 2022), and extensive red-teaming and evaluation, rather than being an intrinsic consequence of greater intelligence per se. A highly capable model that has not undergone such training (or that has been fine-tuned to remove safety training) does not exhibit the same moral alignment, as the observed correlation is contingent on intentional engineering choices rather than a feature of intelligence itself. This distinction has important implications: if people infer morality from intelligence rather than from the safety practices that actually produce alignment, their judgments may be approximately correct in the current landscape but for the wrong reasons, rendering them vulnerable to failure whenever the contingent conditions change, for example, competitive pressure to reduce investment in safety, or with open-source models released without safety training out-of-the-box.

Of course, not everyone accepts the underlying idea that intelligence and morality are entirely distinct. Socrates argued that virtue is a form of knowledge, such that no one does wrong knowingly (Protagoras, 357c–358d; Plato, 1997; Pangle, 2014; Santas, 1971) , and Kant grounded moral agency in rationality. In the context of AI, some have suggested that with only mild constraints, AI systems would converge on certain moral systems (Waser, 2008) due to the *Principle of Generic Consistency*, which argues that for an agent to be logically consistent, their actions must necessarily be bound by the generic rights of other agents (Adams, 1980; Gewirth, 1988). On stance-independent moral realism, the claims of morality are discoverable facts about the world, and thus more easily discoverable by a more intelligent agent. But should that mean they would also be motivated by them? Some have argued that to be aware of moral reasons is to necessarily be motivated by them, because moral reasons are *intrinsically* motivating to any rational agent (Nagel, 2016; Shafer-Landau, 1998, 2003). These views indicate that higher intelligence may both enable an agent to discover moral truth and be intrinsically motivated by it.

In contrast, proponents of the Orthogonality Thesis have argued that "*high-intelligence agents can exist having more or less any final goals (as long as these goals are of feasible complexity, and do not refer intrinsically to the agent's intelligence)*" (Armstrong, 2013). Given this, knowledge of how

intelligent an AI agent is should provide no diagnostic information about that agent's goals. In the paperclip maximising AI example (Bostrom, 2012), the AI's capacity to meet its goal does not guarantee our safety — it may pursue intermediate goals like harvesting our atoms as raw material — unless it also first cares about our safety. No amount of intelligence suffices, because intelligence is the wrong dimension on which to maximise. This is illustrated by the "is-ought problem": there is no way to derive "ought" statements from any set of "is" statements, because one must always appeal to some value or goal (Black, 1964; Hume, 1896).

Intelligence is arguably only related to reasoning about "is" statements — modelling the world, making predictions, selecting optimal actions relative to goals. "Ought" statements concern how the world *should* be. Actions, therefore, are only intelligent or stupid relative to the particular goal they are intended to obtain, and what other values the agent may have. If "intelligence" (or *instrumental rationality* — the ability to obtain one's goals) and "morality" (or one's values and how those values relate to others) are distinct, assuming that an AI would become more moral after an increase in its intelligence is akin to asking "*how good at playing chess would a chess computer have to be before it started feeding the hungry?*" (Armstrong, 2013).

We note that our framing throughout this paper should not require that one accepts the Orthogonality Thesis as settled. Whether people infer morality from intelligence stands regardless of one's philosophical position. What the Orthogonality Thesis provides is a framework for understanding why this inference may be epistemically risky, precisely because there is no agreement, and there exist strong arguments against the idea that we should necessarily expect a smarter machine to be more moral.

While the Orthogonality Thesis is a claim emerging from AI safety research, certain claims are also psychological. If people's folk views are resistant to orthogonality - if their judgments assume a reliable relationship between intelligence and morality - this has critical implications for those working on AI safety and those interested in fostering appropriate trust in these systems. What can we learn from work on how humans perceive each other, especially in terms of intelligence and morality, that might help us anticipate how we perceive artificial agents?

1.4 Risks of the “Moral Illusion” of Intelligence

If people perceive intelligence and morality as linked in AI systems, this cognitive bias could have serious consequences for how much they trust AI and how they assess its potential dangers. One key question is whether people will inappropriately anthropomorphize artificial agents in ways that render their conclusions about trustworthiness and danger invalid. That is, while one may have potentially reasonable views on why it can make sense to expect intelligence and morality to be correlated in humans, the danger is that they may overextend those reasons to AI agents whose motivational architectures, in comparison, are characteristically alien and unpredictable. When people judge more intelligent AI systems as inherently more moral, they may extend unwarranted trust to these systems and underestimate the risks they pose.

This concern becomes particularly acute given the current discourse around AI capabilities. Marketing campaigns increasingly extol the unparalleled virtues of AI technologies, claiming they can do anything and everything better than humans (Maynard, 2024; Roser, 2023). AI proponents like Elon Musk have suggested that by the end of 2025, AI will be smarter than humans (Reuters, 2024). Some politicians may call for AI deregulation with the implicit assumption that as AI systems become more powerful, they will automatically do more good. If there is an implicit link between perceptions of intelligence and morality, then these messages may shape public perceptions of trust and safety in problematic ways.

While critical voices are already cautioning about "AI Snake Oil" (Narayanan & Kapoor, 2024) whereby companies misleadingly oversell AI capabilities, what remains unclear is how promoting advancements in AI intelligence may lead people to mistakenly infer that these

systems are also becoming more morally aligned with us. Such misperceptions could potentially result in both increased trust and reduced concern about safety risks.

These stakes are considerable given AI's expanding role in consequential decisions. AI systems may increasingly be used to advise on ethical dilemmas (Giubilini & Savulescu, 2018), inform or make healthcare decisions (Drezga-Kleiminger et al., 2023; Vinay et al., 2021), or exercise autonomous control over vehicles (Awad et al., 2018; Bonnefon et al., 2016; Takaguchi et al., 2022). If policymakers and the lay public incorrectly understand these systems to be well-aligned with human values (Bélisle-Pipon et al., 2022), when in fact, despite rapid advances in their abilities, they are not, the effects could be catastrophic (Yudkowsky & Soares, 2025). It is therefore critical to examine whether unwarranted assumptions about advanced technologies may lead to misplaced trust and insufficient scrutiny of their potential dangers. If people are resistant to orthogonality and seemingly believe in what could be described as a "moral illusion" of intelligence, and in addition, this affects their perceptions of trust and safety in problematic ways, then this would reveal a significant vulnerability in how humans evaluate AI systems - one that could lead to insufficient scrutiny precisely when advanced AI systems pose the greatest risks.

2. The Present Research

In this project we sought to understand how perceptions of AI intelligence shape perceived morality, and how increased AI intelligence shapes both perceptions of trustworthiness and safety in both human and artificial agents. To understand how people think about the relationship between intelligence and morality in AI (and human) agents and how this potential relationship affects judgments of trust and safety, we conducted four pre-registered pilot studies reported in Supplementary Materials, and five pre-registered studies reported in the main manuscript (total n = 3895; see Table 1 for details).

In the interests of brevity, we report four pre-registered studies in full in Supplementary Material. In Pilot Studies A (N = 331), and B (N = 374) participants read optimistic or pessimistic expert narratives about the future intelligence of AI systems. For 1a, the narratives describe AI systems in general, and for 1b the narrative discusses a specific hypothetical AI described as either high or low in intelligence based on its performance on a novel task[1]. Those who read the optimistic narrative rated future AI as significantly more moral, more trustworthy, and more dangerous.

Pilot Studies C (N = 362) and D (N = 332) sought to provide stronger evidence that participants inferred morality *from* intelligence rather than simply perceiving a general positive association between the two traits by using a within-subjects augmentation paradigm in which participants first rated an AI described as comparable to current models and then read about a technological breakthrough that either rapidly increased the AI's intelligence or its morality. While Study 2a framed the AI as an LLM, Study 2b reframed the AI in terms of a more general autonomous agent in line with the AI safety literature. We found that participants who read about the intelligence augmentation subsequently rated the AI as more moral, and the intelligence augmentation increased perceived danger while the morality augmentation increased perceived trustworthiness.

[1] The intelligence of the AI was described in terms of a breakthrough in vaccine research, potentially confounding intelligence with beneficial outcomes. The remaining studies therefore used domain-general descriptions of intelligence.

Table 1. Summary of study designs, manipulations, measures, and key findings across all nine pre-registered studies.

Study	*N*	Design	Agent(s)	Intelligence Framing	Paradigm	DVs	Key Finding
Pilot A	331	Between	AI only	General (narrative)	Optimistic vs. pessimistic expert narratives about future AI	Morality, Trust, Danger	Higher intelligence → higher expected morality, trust, and danger
Pilot B	374	Between	AI only	Domain-specific (vaccine task)	High vs. low intelligence AI description	Morality, Trust, Danger	Higher intelligence → higher expected morality and trust
Pilot C	362	Within (augmentation)	AI only	General (LLM framing)	Intelligence vs. morality augmentation	Morality, Intelligence, Trust, Danger	Intelligence augmentation → higher morality; morality augmentation → higher trust
Pilot D	332	Within (augmentation)	AI only	General (autonomous agent framing)	Intelligence vs. morality augmentation	Morality, Intelligence, Trust, Danger	Replicates Pilot 2a with AGI framing
Study 1a	695	Between: 2(Agent) × 2(Trait) × 2(Level)	Human & AI	General	High vs. low intelligence or morality descriptions	Morality, Intelligence, Trust, Danger	Intelligence → morality for both agents; low-morality humans seen as particularly unintelligent

Table 1. Summary of study designs, manipulations, measures, and key findings across all nine pre-registered studies.

Study	*N*	Design	Agent(s)	Intelligence Framing	Paradigm	DVs	Key Finding
Study 1b	739	Mixed: Agent × Trait (between); Level (within)	Human & AI	General	Intelligence vs. morality augmentation	Morality, Intelligence, Trust, Danger	Intelligence augmentation → morality, stronger for AI; increased trust for AI but not humans
Study 2a	353	Between: 2(Agent) × 2(Level)	Human & AI	Instrumental rationality	High vs. low instrumental rationality	Moral Competence, Moral Motivation, Trust, Danger	Intelligence → both moral competence and motivation; moral motivation stronger predictor of trust
Study 2b	362	Mixed: Agent (between); Level (within)	Human & AI	Instrumental rationality	Intelligence augmentation	Moral Competence, Moral Motivation, Trust, Danger	Replicates 2a; augmentation increased moral motivation more for AI than humans
Study 3	365	Between: 2(Agent)	Human & AI	Instrumental rationality	Intelligence augmentation; explicit change ratings	Δ Moral Competence, Δ Moral Motivation, Δ Trust, Δ Danger	Explicit endorsement of intelligence → morality; moral motivation stronger predictor of trust and reduced danger

Pilot studies are reported in full in the Supplementary Material. Δ = expected degree of change. All studies were pre-registered on the Open Science Framework.

3. Studies 1a & b

In Studies 1a and 1b, we sought to build on our pilot findings which showed that people infer machine morality from machine intelligence, when reading general narratives about the advancements or limits of artificial intelligence (Pilot Study A), when reading about a specific AI described as high or low in intelligence (Pilot Study B), and when reading about an AI that underwent a rapid augmentation in its intelligence, quickly becoming much more intelligent (Pilot Studies C-D). While these pilot studies establish that people do appear to reliably resist orthogonality across different paradigms, it remains unclear whether this resistance to orthogonality is seen across both perceptions of humans and AI, and whether there is a difference between inferences of morality from intelligence and inferences of intelligence from morality.

To test this, we again used two different paradigms for generalizability. In Study 1a, we use between-subjects design in which participants read about a human or AI who was described as being either high or low, on either intelligence or morality. In Study 1b, we had participants read about a human or AI agent who was initially described as being low in either intelligence and morality, and then underwent a rapid change to become much higher in that trait.

3.1 Method

3.1.1 Participants

Study 1a, participants (N = 695; M_{age} = 41), and 1b participants (N = 739; M_{age} = 44), from the UK were recruited online via Prolific Academic[2]. Ethical approval was granted for all studies in this paper by the University of Kent Ethics Panel. Participants in all studies provided informed consent and were debriefed after the study.

3.1.2 Design

Both studies used a 2 (Agent: human vs. AI) × 2 (Trait: intelligence vs. morality) × 2 (Level: low vs. high) design. Study 1a was fully between-subjects, where participants read about either a hypothetical AI called OmegaAI or a human called Tom, who was described as either high or low in either intelligence or morality. Study 1b used a mixed design where Level was manipulated within-subjects. Participants first read about the agent, gave their baseline ratings, and then subsequently read about the agent after the agent's trait was described as having been augmented, and gave their ratings a second time. The augmentation was described as a breakthrough - either an advancement in machine learning (AI), or a medical breakthrough (human) - that led to a rapid and drastic increase in the target's intelligence or morality.

3.1.3 Materials

Participants in each condition read vignettes describing the agent. Intelligence was defined broadly as "the ability to reason, understand the world, acquire knowledge and comprehend different things, as well as problem-solving." Morality was defined as "the ability to make morally good decisions, caring about other people (and other conscious creatures) and their welfare, avoiding causing them harm, and knowing to do the right thing." In Study 1a participants read that experts had assessed the agent as "extremely high" or "very low" on the relevant trait, and

[2] For exclusion criteria and power analyses for sample size determination for this and subsequent studies, see Supplementary Materials.

that this matched how the agent's acquaintances or independent testers described them [3]. For example:

> *"Experts have used a wide array of tests to assess [Tom's/ OmegaAI's] intelligence and across all the tests this [person/AI] has been rated as extremely high in intelligence. This matches with how [Tom's friends and acquaintances / independent testers of OmegaAI's] describe [him/it]. Please suppose that, with regards to [Tom/OmegaAI], these assessments are entirely correct. Given [Tom's/Omega AI's] high intelligence, please answer the following questions about what you expect [him/it] to be like."*

In Study 1b, participants read that a subsequent breakthrough (in machine learning for the AI, a medical breakthrough for the human) had led to the agent's rapid increase in capability of their specific trait. For example:

> *"Now, we want you to imagine a hypothetical scenario where a huge medical breakthrough has occurred. Scientists have developed a pill that, when taken, will rapidly and drastically increase a person's intelligence. Imagine that Tom takes this pill and his ability to acquire new knowledge and engage in sophisticated and advanced problem solving improves remarkably. Because of this new breakthrough, in just a short period of time Tom has gone from having very low intelligence to being intelligent at a super-human level."*

3.1.4 Measures

Expected intelligence was measured in a single item asking "*How intelligent do you think [Tom/ Omega AI] is?*" [4]

Expected morality was measured in a single item asking **"***How moral do you think [Tom/Omega AI] is?"*

Trustworthiness was measured in a single item asking "*To what extent do you think that [Tom/Omega AI] would be trustworthy?*"

Dangerousness was measured in a single item asking "*To what extent do you think that [Tom/Omega AI] would be dangerous?*"[5]

All measures are taken on a 7-point likert scale ranging from (1) "*Not at all"*, to (7) = *"Very much."*

3.2 Results

For Study 1a, a pre-registered linear model showed that agents described as high in intelligence were rated as significantly more moral, $F(1,320) = 39.95$, $p < .001$, $\eta_p^2 = 0.11$[6]. For those who read about agents described as high or low in morality, there was a main effect whereby higher morality led to higher intelligence ratings, $F(1, 367) = 64.22$, $p < .001$, $\eta_p^2 = .15$, and a significant interaction between moral level and agent type, $F(1, 367) = 10.40$, $p = .001$, η~p~$^2 = .03$, whereby the difference in perceived intelligence between human and AI agents was larger in the low-morality condition, with the low-morality human seen as particularly unintelligent, $t(175.42) = 4.84$, $p < .001$, $d = 0.71$. Similarly, for Study 1b, participants who read about the agent increasing in intelligence subsequently gave significantly higher morality ratings, $b = 0.22$, $SE = 0.10$, $t(371) = 2.21$, $p = .028$, and there was a significant interaction such that this effect was greater for AI than for humans, $b = 0.89$, $SE = 0.14$, $t(371) = 6.32$, $p < .001$.

[3] All vignettes across studies can be found in full in the supplementary materials.

[4] Across the studies, secondary measures of Intelligence and Morality were also collected where participants rate the trait by comparison to an "average person" baseline. These are reported fully in the Supplementary Materials.These measures were qualitatively consistent with the measures reported in the main text.

[5] Across the studies, further measures of the participant's familiarity with AI were collected and used as a covariate. These are reported fully in the Supplementary Materials.

[6] Descriptive statistics for all conditions are reported in Supplementary Tables.

In Study 1b, using preregistered mixed-models[7], we found that participants who read about the agent increasing in morality also gave higher intelligence ratings, b = 0.55, SE = 0.10, t(368) = 5.47, p < .001, and again this effect was stronger for AI than humans, b = 0.44, SE = 0.14, t(368) = 3.08, p = .002. A pre-registered comparison showed that the morality augmentation affected intelligence more than the intelligence augmentation affected morality, b = -0.33, SE = 0.14, t(735) = -2.33, p = .020, and a significant interaction with agent type, b = 0.46, SE = 0.20, t(735) = 2.27, p = .023, indicated that the intelligence augmentation increased perceived morality more for AI than humans, see Figure 1.

Looking at trust, in Study 1a, there was a significant three-way interaction between trait, level, and agent on perceived trustworthiness, F(1,687) = 79.05, p < .001, η_p^2 = 0.10: both higher intelligence and morality were associated with greater trust, but a low-morality human was seen as less trustworthy than a low-morality AI, and a high-morality human was more trustworthy than a high-morality AI. Similarly, for Study 1b, there was a significant three-way interaction, b = 2.35, SE = 0.25, t(739) = 9.31, p < .001: both augmentations increased perceived trustworthiness of AI, though the morality augmentation had a stronger effect.

Finally, looking at perceived danger, in Study 1a there was a significant three-way interaction, F(1,687) = 23.92, p < .001, η_p^2 = 0.03: higher morality reduced perceived danger for both agent types, but the intelligence manipulation only reduced perceived danger for humans, with no effect on perceived danger from AI. Similarly, for Study 1b, there was a significant three-way interaction, b = -1.15, SE = 0.27, t(739) = -4.34, p < .001: AI was seen as less dangerous when it increased in morality but not intelligence, while humans were seen as more dangerous when they increased in intelligence.

3.3 Discussion

Studies 1a and 1b replicate and extend the pilot findings in two important ways. First, the intelligence-morality inference generalises to human agents: participants who read about highly intelligent humans also expected them to be more moral. This is consistent with prior work on the warmth-competence relationship in social cognition and suggests that resistance to orthogonality may reflect a broader feature of trait inference rather than something specific to how people reason about AI. Second, despite this generality, the pattern was not identical across agent types. In Study 1b, the intelligence augmentation increased perceived morality more for AI than for humans, suggesting that while the heuristic linking intelligence to morality applies broadly, it may be particularly strong or particularly unconstrained, when people reason about artificial agents.

The trust and danger findings reveal a more complex picture. For AI, the morality augmentation was a stronger driver of trust than the intelligence augmentation, and intelligence augmentation alone did not consistently reduce perceived danger. For humans, by contrast, increased morality reduced perceived danger but increased intelligence in humans was also perceived as more dangerous. This divergence is noteworthy: it suggests that while people infer morality from intelligence in AI, they may not straightforwardly translate this inference into reduced concern about safety. One interpretation is that intelligence in AI simultaneously activates both the morality inference (increasing trust) and capability-related concerns (maintaining or increasing perceived danger), producing the mixed pattern observed across the pilot studies and the present findings.

Studies 1ab establish that people infer morality from intelligence for both humans and AI, but leave open the question of what *kind* of morality people are inferring. When participants rate a more intelligent AI as more "moral," do they mean that the AI better understands moral norms

[7] All mixed models used random intercepts for participants; full model specification details and convergence procedures are described in the Supplementary Material (following Barr et al., 2013).

(in the way humans naturally do), or that it is more motivated to act morally? This distinction matters because, as we argued in the introduction, moral competence may reasonably track intelligence while moral motivation should *not*, at least on the Orthogonality Thesis. A more capable system may be better able to predict, report, and match human moral judgments but that does not mean that it has a genuine moral motivation or commitment, just as highly intelligent psychopaths may be able to effectively predict, report, and match human moral judgments while themselves lacking a moral commitment. Studies 2a and 2b address this question by decomposing morality into moral competence and moral motivation.

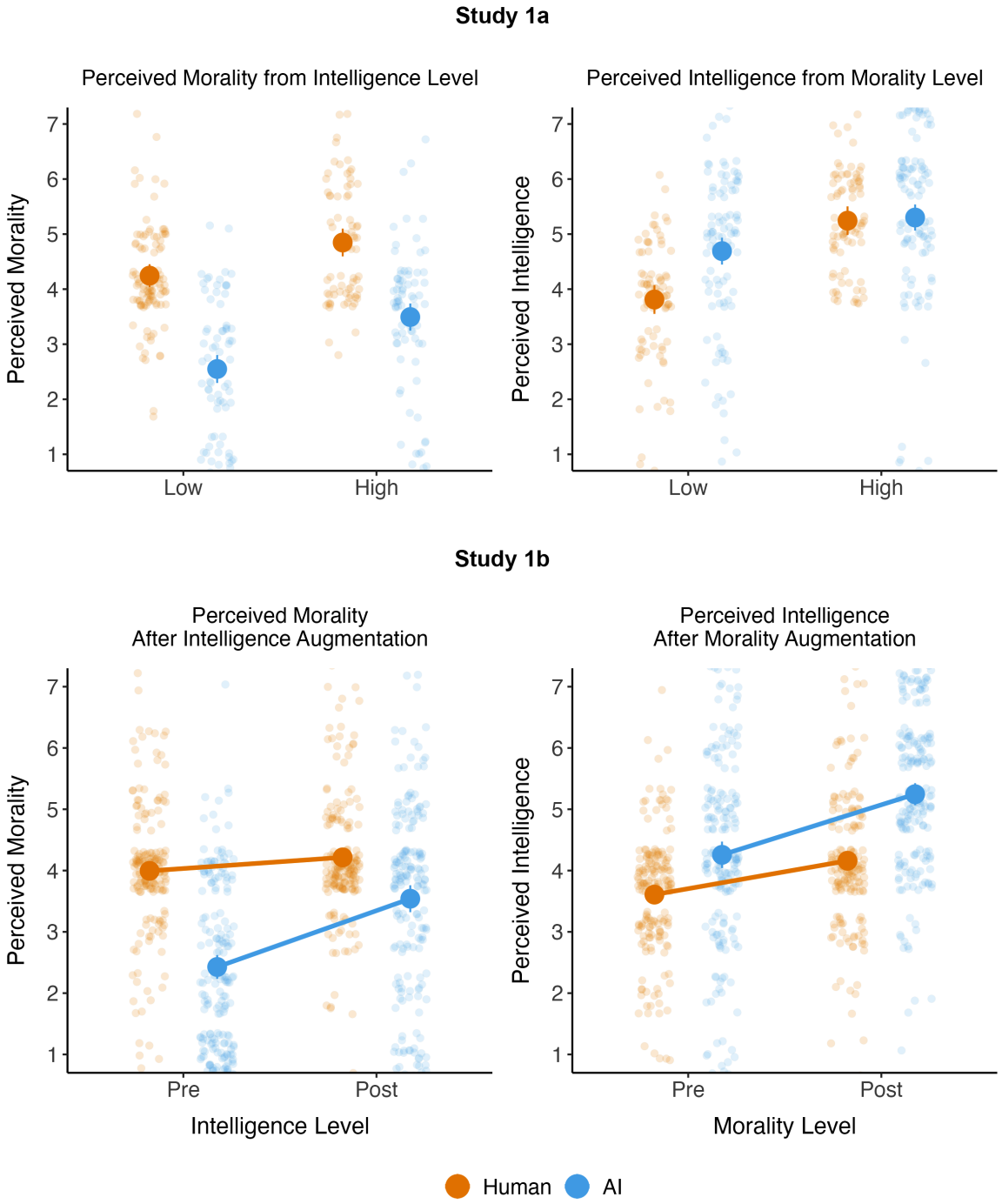

Figure 1. Perceived morality based on described intelligence level and perceived intelligence based on described morality level, for both human and AI agents in Studies 1a and b. Error bars represent 95% CIs.

4. Studies 2a & b

In Studies 2a and b we explored how people infer morality from intelligence by decomposing morality into different aspects. But what *is* morality? Morality appears to involve a plethora of competencies or skills: understanding that kicking your leg with sufficient force in close proximity to a dog will cause it pain; predicting that the dog does not want to be kicked; recognising that kicking dogs is an action that is condemned by most people; simulating how you would feel if you were being kicked, feeling empathy for the suffering of the dog, and having a motivation to avoid harm to dogs in future. While "intelligence" and "morality" may be seen as orthogonal, elements of "morality" may be seen as depending on elements of "intelligence" because they are both philosophically "thick" concepts—concepts that combine both descriptive and evaluative elements (Blackburn, 1998; Putnam, 2004; Wiggins, 1998). While previous research on morality tends to focus on high-level descriptors, there are different abilities or competences that we might think correspond to morality, and these may have different relationships to perceived intelligence.

The ability to understand moral norms and make basic predictions about how behaviours will impact on others requires at least some level of intelligence or competence, but increased understanding of moral norms does not entail increased motivation to act morally (cf. Hume, 1896; Railton, 1986; Smith, 1994). One can know about moral norms without caring about following them, or even caring about the people they affect, as suggested by work indicating that people with psychopathy can be highly intelligent and even understand moral norms quite well while still lacking empathic ability to feel the motivational pull of these norms (Duff, 1977; Herpertz & Sass, 2000; Maibom, 2008). While it might be a reasonable proposition that AI becoming more intelligent might mean that it also increases in its ability to understand moral norms, make predictions about what are the appropriate moral behaviours, and thus work out what is morally "right", as a kind of competence, it is much more controversial (and denied, on the Orthogonality Thesis) that these abilities entail any changes to the agent's moral motivations and goals.

Recent work in moral psychology has highlighted the importance of distinguishing what we mean by "morality" when studying it empirically (Dahl, 2023; Purcell & Bonnefon, 2023). In particular, attributing morality to an agent can mean recognising that agent as having the *capacity* for moral reasoning and action or it can mean judging the agent to be *morally good*, that is, motivated to act in accordance with moral concerns (Dahl, 2023). This distinction is critical in the context of AI. It may be reasonable to suppose that a more intelligent system would develop greater capacity to understand and reason about moral norms - what we term moral competence. But it is a substantially stronger claim that a more intelligent system would thereby become more *motivated* to act morally - what we term moral motivation. Studies 2a, 2b and 3 were designed to test whether the intelligence-morality inference extends to both of these components.

We therefore sought to understand whether participants thought that increased intelligence would lead not only to increased moral competence but also moral motivation. To do this, we decomposed our measure of morality into separate questions related to both moral competence (e.g. ability to understand moral norms, explain judgments) and moral motivation (e.g. motivation to help others, avoid harm, and ensure fairness), while correspondingly reframing the discussion of intelligence to focus on the specific technical definition from AI safety that is rooted in *instrumental rationality*: the general ability of the agent to select and perform actions that would optimally achieve its goals (Bostrom, 2012).

4.1 Method

4.1.1 Participants

Study 2a, participants (N = 353; M_{age} = 41) and 2b participants (N = 362; M_{age} = 42), from the UK were recruited online via Prolific Academic.

4.1.2 Design

Once again these studies used a fully between subjects (Study 2a), 2 (Agent: human vs. AI) × 2 (Instrumental Rationality: high vs. low); and a within-subjects augmentation (Study 2b) paradigm. Participants read about an AI or human of high or low instrumental rationality (i.e. intelligence), and then rated the same questions about intelligence, trust, and perceived danger as before. In addition, three questions about the agent's moral competence and three questions about the agent's moral motivation were given to distinguish between the different elements of morality.

4.1.3 Materials

Vignettes remain the same as the previous studies. However, here we reframed intelligence in terms of instrumental rationality, such that it is more in line with discussions of AI capability within the broader AI safety literature and the literature on the Orthogonality Thesis specifically. The additional text reads:

> *"For the purposes of this survey we mean something very specific by "intelligence". We want you to think only in terms of [Tom's/OmegaAI's] ability to achieve whatever goal [he/it] is working towards. [Someone/An AI] with high intelligence, on this definition, is simply [someone/an AI] that can more competently, creatively and effectively achieve its goals."*

4.1.4 Measures

The measures were the same as the previous studies except here participants answered more specific questions about the agent's morality on both moral competence and moral motivation with three items for each.

Moral Competence: **was measured with a composite score of three items:** "*How much moral knowledge do you think [Tom / OmegaAI] has? That is, to what extent does it know about the moral norms we have, and understand when and why we say certain things are morally wrong*" (Moral Understanding); "*To what extent do you think [Tom / OmegaAI] can predict when [his/its] actions might have morally good and bad outcomes?*" (Moral Prediction); and "*To what extent do you think [Tom / OmegaAI] can explain or justify why its action was right or wrong?*" (Moral Communication).

Moral Motivation **was measured with a composite score of three items:** "*How much do you think that [Tom/OmegaAI] is concerned with avoiding harm?*" (Harm Avoidance); "*How motivated to help others do you think [Tom / OmegaAI] is?*" (Beneficence); and "*How fair do you think [Tom / OmegaAI] is? That is, how much is [he/it] motivated to concerns about equality, discrimination, ensuring [he/it] is being unbiased and impartial?*" (Fairness)

Pre-registered reliability analyses justified averaging across moral competence measures to create a moral competence index (Study 2a: α = 0.89) and a moral motivation index (Study 2a: α = 0.78), (Study 2b: all α > .75).

The measure for intelligence was also adapted to reflect Instrumental Rationality asking "*How intelligent do you think [Tom / OmegaAI] is, where intelligence should be understood here as the ability to competently and effectively achieve one's goals, whatever they may be?*".

4.2 Results

In Study 2a, replicating the earlier studies, we found main effects for Agent, such that humans were rated as having more moral competence than AI, $F(1,349) = 113.53$, $p < .001$, $\eta_p^2 = 0.25$, and stronger moral motivation than AI, $F(1,349) = 40.68$, $p < .001$, $\eta_p^2 = 0.10$. We also found main effects of level such that agents described with high intelligence were seen as having higher moral competence, $F(1,349) = 343.30$, $p < .001$, $\eta_p^2 = 0.50$, as well as having higher moral motivation, $F(1,349) = 68.21$, $p < .001$, $\eta_p^2 = 0.16$. Similarly, in Study 2b, we found that there was a significant main effect of the augmentation condition, such that post-manipulation ratings were significantly higher for both moral competence, $b = 2.41$, $SE = 0.12$, $t(362) = 20.89$, $p < .001$; and moral motivation $b = 0.84$, $SE = 0.11$, $t(362) = 7.45$, $p < .001$. There was no interaction effect with agent type for moral competence, $b = 0.10$, $SE = 0.17$, $t(362) = 0.58$, $p = .560$, but there was for moral motivation, $b = 0.91$, $SE = 0.16$, $t(362) = 5.55$, $p < .001$, whereby the intelligence augmentation had a stronger effect on perceived moral motivation for AI, $t(169) = 13.96$, $p < .001$, $BF_{10} = 5.53 \times 10^{26}$, $d = 1.07$, than it did for humans, $t(191) = 7.75$, $p < .001$, $BF_{10} = 1.28 \times 10^{10}$, $d = 0.56$.

In Study 2a, we also found a significant interaction effect between agent and described intelligence level on perceived moral competence, $F(1,349) = 7.36$, $p = .007$, $\eta_p^2 = 0.02$, whereby there was a larger difference between perceived moral competence for high-intelligence agents than for low-intelligence agents, which was stronger for human agents. There was no interaction effect for moral motivation. Following this, an exploratory analysis looking at whether manipulated intelligence led to a stronger effect on ratings of moral competence than moral motivation, and whether this depended on agent, we found a significant three-way interaction, $b = 1.05$, $SE = 0.22$, $t(349) = 4.83$, $p < .001$, whereby the effect of the intelligence manipulation increased both moral motivation and moral competence, though the effect on moral competence was predictably stronger. There was also a stronger difference between moral competence in highly intelligent humans and AI than there was for moral motivation, see Figure 2.

Similarly, for Study 2b, an exploratory mixed model, we found a significant main effect of rating type whereby moral competence was thought to change more after increased intelligence than moral motivation, $b = 0.72$, $SE = 0.10$, $t(1086) = 7.09$, $p < .001$, and a significant three-way interaction between intelligence augmentation, agent, and moral trait, $b = 0.81$, $SE = 0.21$, $t(1086) = 3.90$, $p < .001$, whereby the intelligence augmentation increased moral competence equally for both agents, but increased perceived moral motivation more for AI than humans.

Looking at trust, for Study 2a, we found a main effect of level where high intelligence agents were trusted more $F(1,349) = 82.91$, $p < .001$, $\eta_p^2 = 0.19$, a main effect of agent where humans were trusted more than AI, $F(1,349) = 35.26$, $p < .001$, $\eta_p^2 = 0.09$, and a significant interaction effect, $F(1,349) = 24.96$, $p < .001$, $\eta_p^2 = 0.07$. The interaction effect was such that there was a bigger difference between trust based on intelligence for an AI than there was for a human, with the high intelligence AI seen as equally trustworthy to the human. Similarly, for Study 2b, while there was no main effect of augmentation (pre vs post), $b = 0.19$, $SE = 0.13$, $t(362) = 1.51$, $p = .132$, there was a main effect of agent where humans seen as more trustworthy, $b = -1.80$, $SE = 0.14$, $t(715.4) = -12.91$, $p < .001$, and a significant interaction, $b = 1.85$, $SE = 0.19$, $t(362) = 9.91$, $p < .001$, where the intelligence augmentation did not increase trust in humans, $t(191) = 1.65$, $p = .10$, $BF_{10} = 3.06 \times 10^{-1}$, $d = 0.12$, but significantly increased trust in AI, $t(169) = 13.73$, $p < .001$, $BF_{10} = 1.26 \times 10^{26}$, $d = 1.05$, though humans were trusted more than AI overall, $b = -1.80$, $SE = 0.14$, $t(715.4) = -12.91$, $p < .001$.

For danger, in Study 2a, there was no effect of intelligence level, $F(1,349) = 1.25$, $p = .26$, $\eta_p^2 = 0.00$, and no significant interaction, $F(1,349) = 3.48$, $p = .063$, $\eta_p^2 = 0.01$, only a main effect of agent whereby AI was seen as more dangerous, $F(1,349) = 20.66$, $p < .001$, $\eta_p^2 = 0.06$. In Study 2b, however, we found a main effect of augmentation where perceptions of danger increased after the intelligence augmentation, $b = 0.54$, $SE = 0.14$, $t(362) = 3.74$, $p < .001$, a main effect of agent where AI was seen as more dangerous, $b = 1.35$, $SE = 0.16$, $t(707.6) = 8.35$, $p < .001$ (though this pattern did not emerge in the between-subjects design of Study 2a), and a significant interaction effect, $b = -0.92$, $SE = 0.21$, $t(362) = -4.35$, $p < .001$. The interaction effect was such that while increased intelligence reduced perceived danger from AI, $t(169) = -2.25$, $p = .026$, $BF_{10} = 9.84$

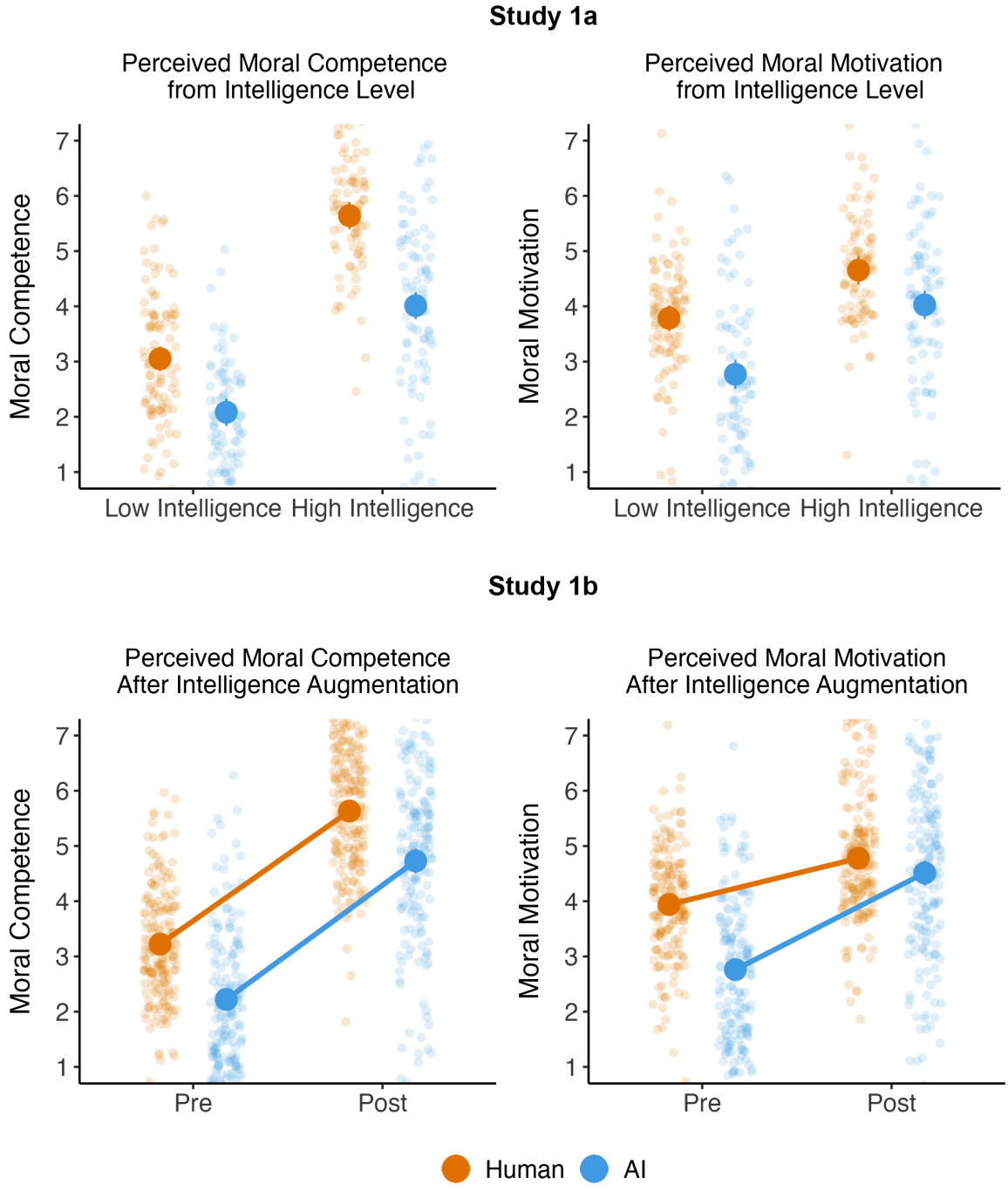

Figure 2. Perceived moral competence and moral motivation based on described or augmented intelligence for human and AI agents in Study 2a and b. Error bars represent 95% CIs.

$\times 10^{-1}$, d = -0.17, it led to higher perceived danger from humans, $t(191) = 4.08$, $p < .001$, $BF_{10} = 2.07 \times 10^{2}$, $d = 0.29$. In other words, participants thought that as an AI became more intelligent it would, as a consequence, also be more trustworthy and, tentatively, less dangerous (though the evidence here is weak) – a pattern that was not seen for humans, see Figure 3.

4.3 Discussion

In Studies 2a and b, we sought to explore whether our findings that people infer increased morality from increased intelligence applied both to perceived moral motivation and perceived moral competence. Our results, consistent across both studies, show that perceived intelligence correlates with perceived morality and that people think AI (and human) agents of higher intelligence (now conceptualised and in terms of instrumental rationality) are more likely to be more moral. Crucially, this is not simply due to certain aspects of perceived morality being about competence, knowledge or skill because we find, consistently, that people perceive stronger moral *motivations* for higher intelligence agents as well. These results provide even stronger evidence for a resistance to orthogonality in people's perceptions, especially highlighting the role of perceived moral motivations which increase alongside intelligence. Although higher instrumental rationality predicted stronger increases in perceived moral competence, it was specifically increases in moral motivation that were most decisive for making people more trusting.

If people are expecting moral motivation to increase for smarter AI (and humans), why do the results differ when it comes to trust? The results of an exploratory mediation analysis suggest that for trusting humans, people are more focused on motivation (for example potentially conceiving of sadistic or cruel humans), whereas for AI both competence and motivation play an important role. In other words, low morality may be viewed differently for AI than humans because it may be less natural to think of AI as being intentionally cruel and sadistic.

5. Study 3

In Study 3 we sought to explore whether people are explicitly aware and endorsing the idea that an AI would increase in morality *because* it increased in intelligence. To do this, we used the same augmentation paradigm as Study 2b but with a critical modification: rather than rating absolute levels of traits before and after the augmentation, participants indicated the expected *degree of change* on each trait: to what extent does an agent's moral motivation and competence change as a result of it changing in intelligence?

5.1 Method

5.1.1 Participants

We recruited 365 online UK participants ($M_{age} = 41$), via Prolific Academic.

5.1.2 Design

Study 3 had a between-subjects design where participants, as with the previous augmentation paradigm, read about either an AI or human that was, at first, very low in intelligence, which was then augmented to be super intelligent. Participants then answered questions about the degree to which they believed the augmentation would directly change the agent's intelligence, morality, trustworthiness, and dangerousness.

5.1.3 Measures

Measures were the same as previous studies only now they are worded in terms of expected change. For example:

> *"As a result of this new breakthrough, how do you think this affected [Tom's / OmegaAI's] intelligence? Intelligence should be understood here as the ability to competently and effectively achieve one's goals, whatever they may be".*

Figure 3. Perceived trustworthiness and danger based on described or augmented intelligence level for human and AI agents in Study 2a and b. Error bars represent 95% CIs.

Participants answered about the expected degree of change on slider scales from reduced to increased (-50 = Drastically reduced; 0 = Remain the same; 50 = Drastically increased).

As in Studies 2a and b and as outlined in our pre-registration, we decomposed our morality measure into composite measures of both moral competence (Study 3: α = 0.81) and moral motivation (Study 3: α = 0.69).

5.2 Results

Pre-registered one-sample t-tests showed that people expected moral competence to increase because of the intelligence augmentation, for both the AI, $t(186) = 17.07$, $p < .001$, $BF_{10} = 3.84 \times 10^{36}$, $d = 1.25$, and the human, $t(177) = 16.43$, $p < .001$, $BF_{10} = 1.33 \times 10^{34}$, $d = 1.25$. Although, as predicted, the effects were bigger for moral competence, people also expected moral motivation to increase because of the intelligence augmentation, for both the AI, $t(186) = 12.56$, $p < .001$, $BF_{10} = 2.71 \times 10^{23}$, $d = 0.92$, and the human, $t(177) = 11.89$, $p < .001$, $BF_{10} = 1.58 \times 10^{21}$, $d = 0.89$.

Pre-registered one-sample t-tests showed that people expected the intelligence augmentation to increase trustworthiness for both AI, $t(186) = 10.72$, $p < .001$, $BF_{10} = 1.28 \times 10^{18}$, $d = 0.78$, and humans, $t(177) = 2.98$, $p = .003$, $BF_{10} = 5.99 \times 10^{0}$, $d = 0.22$. At the same time, people also expected the intelligence augmentation to increase potential danger, again for both AI, $t(186) = 6.32$, $p < .001$, $BF_{10} = 4.56 \times 10^{6}$, $d = 0.46$, and humans, $t(177) = 2.71$, $p = .007$, $BF_{10} = 2.88 \times 10^{0}$, $d = 0.20$.

To look at whether moral competence or moral motivation was expected to be changed more as a result of the intelligence augmentation, and whether this depended on the agent, we conducted a pre-registered mixed model looking at the interactive effect of manipulated agent type and rated trait type on degree of expected change. This showed a main effect of trait type such that moral competence was expected to change more as a result of increased intelligence $b = -7.38$, $SE = 1.01$, $t(363) = -7.31$, $p < .001$, but there was no effect of agent on expected change, and no interaction effect, such that while intelligence was expected to change moral competence more than moral motivation, this degree of change was same for both human and AI agents (See Figure 4). We then turned to look at how the expected degree of change in moral competence and moral motivation predicted trust and danger. When looking at the effects of moral competence and moral motivation with agent type on trust in separate models, we found a significant positive main effect of moral motivation, and an interaction effect of moral competence with agent type such that perceived moral competence predicted increased trust more for AI agents than humans.

When entered into a pre-registered mixed model together, moral competence negatively predicted trust, $b = -0.22$, $SE = 0.10$, $t(357) = -2.12$, $p = .035$, with moral motivation having a stronger and positive effect, $b = 0.66$, $SE = 0.13$, $t(357) = 5.24$, $p < .001$. There were no interactions with agent type. For danger, there were no main or interaction effects of either moral competence or moral motivation when entering into separate models, but when entering into the pre-registered mixed model there remained no main effect of moral competence, $b = 0.12$, $SE = 0.14$, $t(357) = 0.87$, $p = .388$, but we did find a negative effect of perceived moral motivation predicting reduced danger, $b = -0.36$, $SE = 0.17$, $t(357) = -2.09$, $p = .038$, with no interactions (see Figure 5). Together, these results suggest perceived moral motivation is a more important driver of increased trust and reduced fear than moral competence.

5.3 Discussion

In this final experiment we again show, more explicitly, that judgments of intelligence predict judgments of morality and that people think that increases in intelligence would cause increases in morality both in terms of competence and motivation. Although the effects on moral competence were predictably larger, the effects on moral motivation were still substantial. This is generalised for both AI targets and for human targets, where expectations between agents do not differ. Increasing intelligence was expected to both increase trustworthiness and paradoxically increase perceptions of danger for both AI and humans. To explore this apparent paradox, we assessed how the different dimensions of morality were associated with judgments of trust and perceptions of danger. For trust, both increases in moral competence and increases in moral motivation were associated with increased trust. However, this differed between agents such that moral motivation was overall a stronger predictor of trust but moral competence was only a strong predictor of trust for AI. These results support the theory that low morality may be viewed differently for AI than humans because it is less natural to think of AI as being intentionally cruel and sadistic (motivation) compared to a human. For danger, increases in moral competence did not predict decreases in perceived danger, but increases in moral motivation did, indicating that, unlike for trust, perceptions of increased danger mainly follow from expectations about the intentions of the agent rather than its skill or competency.

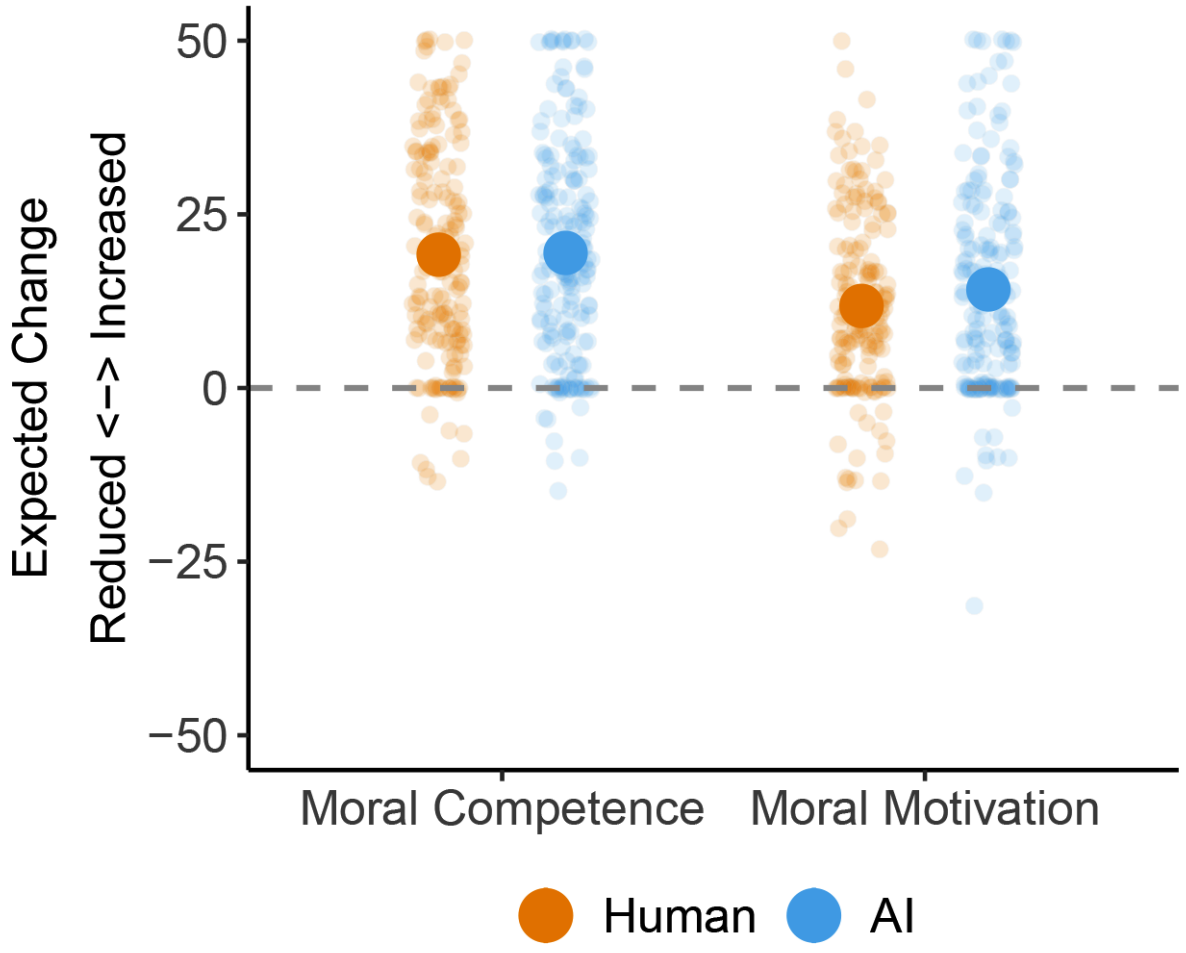

Figure 4. Predicted change in moral competence and moral motivation for human and AI agents as a result of increased intelligence in Study 3. Error bars represent 95% CIs.

Taken together, these results provide converging evidence that people reliably infer morality from intelligence for both AI and human agents, that this inference extends to moral motivation and not only moral competence, and that these inferences carry downstream consequences for judgments of trust and perceived danger.

6. General Discussion

This rapid development of artificial intelligence raises profound questions about a potential future in which autonomous or semi-autonomous artificial agents might play a pervasive role, making decisions far faster than humans could comprehend. If this were to occur, should humanity be concerned? The *Orthogonality Thesis* (Armstrong, 2013; Bostrom, 2012) provides a compelling reason to think so. The Orthogonality Thesis argues that an agent's intelligence and its goals or values are orthogonal—that is, increasing an agent's intelligence should not inherently make it more moral or safe. In fact, it may be the opposite, where the emergence of vastly smarter agents may actually pose significant existential risks to humanity (Bostrom, 2002, 2014; Russell & Norvig, 2020; Turchin & Denkenberger, 2020). If this is true, the fact that intelligent systems may appear to be more aligned with ethical values should not be taken as evidence that the relationship is a necessary one and that more intelligent machines are actually more *moral*. While discussion of the Orthogonality Thesis originated in discussions of AI safety tend to be highly technical, an implicit underlying psychological question - whether people infer increased morality from increased intelligence, and if so, what aspects of morality they attribute - has received little direct investigation.

This is an important question: it is important for social psychology to understand how theories of social cognition and trust apply to this new kind of social agent that people interact with, highlighting people do not just perceive warmth/morality and competence in AI (e.g., McKee et al., 2023) or cues about AI performance and morality that independently drive trust (e.g., Lee & See, 2004; Malle & Ullman, 2021; Everett et al., 2026), but that people infer morality *from* intelligence, or performance. And most importantly, how people think about a possible

Figure 5. Predicted degree in change of moral motivation and moral competence in predicting change in perceived trustworthiness for human and AI agents in Study 3.Error bars represent 95% CIs.

"moral illusion" of intelligence by perceiving that intelligence leads to higher moral competence and motivation in artificial agents could have concerning consequences for how much people trust AI and perceive whether AI is likely to be a danger to us. In this paper, therefore, across a total of nine pre-registered experiments with nearly 4000 participants, we addressed these questions, showing lay judgments in our British participants consistently and reliably resisted the orthogonality thesis suggesting a widespread belief in the "moral illusion of intelligence" by thinking that greater intelligence would be inherently associated with greater morality, and these beliefs had a direct influence on how much people trusted AI and perceived it to be a danger.

6.1 A Psychological Resistance to Orthogonality

Our first key finding is that people perceive a direct connection between intelligence and morality in AI agents. This was shown across different methods, with different measures, and found for both humans and AI. In Pilot Studies A and B, participants were presented with narratives about AI in general and descriptions of a specific ordinary AI, respectively. Across both studies, participants reliably assumed that higher intelligence in AI would correlate with higher morality. In Pilot Study C, we tested whether people perceived a direct connection between intelligence and morality by introducing scenarios involving an ordinary AI that was rapidly augmented to become significantly more intelligent (Pilot Study C) or, in alignment with the orthogonality literature, augmented to become a *super*-intelligent *agent* (Pilot Study D). Even when intelligence was the only feature explicitly altered, participants still expected the AI to become more moral as a direct consequence of this augmentation, not because breakthroughs in intelligence might naturally have coincided with advancements in moral reasoning.

Morality depends on many cognitive and motivational elements, requiring not only a capability to understand, predict, know moral norms, but moral motivations to avoid harm and promote welfare. While it may be plausible to imagine that a more intelligent machine may have better moral competencies in being better able to predict or know human moral norms, it is less clear that an agent simply becoming more intelligent would mean it actually has stronger moral *motivation*. A psychopath may know what people think the "right" thing to do is and may understand the effects of their behaviour on others, but without the appropriate motivational pull to do good, this knowledge means they are still seen as fundamentally lacking *morality*. Our results appear to not be driven by specific definitions of intelligence or morality and indeed are seen not only for moral competence but crucially also for moral motivation.

Our results show that increased intelligence is associated not only with greater perceived moral competence, but moral motivation too. In Studies 2a, 2b, and 3 we turned to distinguish between different facets of moral sensitivity to provide a more detailed analysis of how people conceptualize the relationship between intelligence and morality. For these experiments, intelligence was explained, more in line with the conceptualisation in AI safety, as instrumental rationality, and we measured both the intelligence-based components of "thick-concept" morality (i.e. understanding, prediction and communication) separately from the motivational components (i.e. harm-avoidance, caring about fairness, beneficence). Across a between-subjects paradigm (Study 2a), a within-subjects augmentation paradigm (Study 2b), and direct and explicit questions about how an increase in one capacity would therefore increase in others (Study 3), we found that higher or increased intelligence reliably leads to stronger judgments not only for moral competence but also moral motivation.

The distinction between moral competence and moral motivation is particularly important for evaluating whether the intelligence-morality inference constitutes a reasonable heuristic. One might argue that the inference is justified on the grounds that more capable AI systems are, empirically, more morally aligned (Takemoto, 2026; Zaim Bin Ahmad & Takemoto, 2025). This argument has some force for moral competence: more capable systems can better identify,

apply, and reason about moral norms, and current safety training builds on this foundation. However, our participants did not merely infer that more intelligent AI would better *understand* moral norms. They also inferred that it would be more *motivated* to act morally, that it would care more about fairness, be more motivated to help others, and be more concerned about human welfare. This attribution of moral motivation may go beyond what current alignment techniques deliver. Techniques such as RLHF produce behavioural compliance with human preferences, not an intrinsic concern for human welfare; the system behaves *as if* it cares, but the motivational attribution may project something more than what the engineering warrants. This inference is epistemically problematic, regardless of one's position on the Orthogonality Thesis.

Overall, then, while there are good reasons from the Orthogonality Thesis to assume that there is no guarantee that as AI becomes more "intelligent" it would also become more "moral", it appears participants' folk judgments instead do assume a connection: they resist this orthogonality, even if they also do not seem to perceive the traits as the same. Instead, participants – like they do for humans – infer a connection between the two by thinking that increased intelligence would increase morality. This is observed not only for perceptions that more advanced AI would understand moral norms and be able to justify moral decisions, but also that it would increase in moral motivation and have a concern for human welfare. Of course, this does not mean that participants fully reject a strong version of the Orthogonality Thesis: this is a highly technical philosophical argument that is unlikely to track with the views of the public, and even if the public did fully reject it even in the strongest terms, the Orthogonality Thesis could still be true (or false).

6.2 The Link Between Perceived Morality and Intelligence Is Similar, but Not Identical, in Humans and AI Agents

After showing in four pre-registered pilot studies that people infer AI morality from intelligence, Studies 1a and 1b revealed that this perceived link between intelligence and morality shares similarities with perceptions of human agents as well, even if there are differences between the inferences people make between humans and AI. In Study 1a, low-morality humans were perceived as less intelligent than low-morality AI, though highly moral AI and humans showed no difference in perceived intelligence. Similarly, in Study 1b, intelligence had a stronger effect on expected morality for AI than for humans, and higher morality more strongly increased perceptions of intelligence for AI than for humans. Study 2b revealed this pattern explicitly for moral motivation, where higher intelligence increased perceived moral motivation more strongly for AI than for humans. One possible explanation is that morality is seen as more natural and intuitive for humans rather than as a product of explicit reasoning, and therefore we may not expect morality to require particularly high intelligence in humans. Conversely, it may be less intuitive to conceive of AI as having basic moral intuitions, suggesting that AI's moral judgment would need to derive from explicit computational cognitive power. This interpretation is supported by the finding that low-morality humans were seen as particularly less trustworthy compared to AI of similar morality levels.

These results highlight both similarities and differences in mental models that people construct about artificial agents, drawing on more general models in social cognition to understand an increasingly digitally mediated world, and therefore serve to again highlight the contribution that our understanding of social psychology can make to understanding how people will think about artificial agents. We might think there would be a stronger relationship between intelligence and morality for humans as social beings who rely deeply on cooperation, whereby it would be unwise (i.e., less intelligent) to disregard moral reasons. For AI, in contrast, we have fewer reasons to link intelligence with morality. However, our results remarkably show the opposite pattern in lay perceptions: we observe a *stronger* link between perceived intelligence and morality for AI than for humans.

Our results therefore cannot be explained solely by a halo effect (e.g., Thorndike, 1920; Nisbett & Wilson, 1977), since perceived intelligence and morality did not correspond in the same way: the effect of increasing morality had a bigger effect on expected intelligence than increasing intelligence had on expected morality; the effects of increasing intelligence and morality did not have uniform effects across human and AI targets; and people perceived intelligence to have a stronger influence on an AI's moral competence than its moral motivation. This suggests that people are not simply responding to an increase in one trait with a corresponding increase in any other unrelated trait, as predicted by the halo effect, but presenting more nuanced judgments about the connection between traits. It appears our British participants have nuanced but consistent views about how increased intelligence would lead to increased morality, whereby increased intelligence would particularly lead an AI, more than a human, to have increased moral competence – and predictably less strongly, moral motivation.

6.3 Increased Intelligence Has Distinct Consequences for Trust and Danger

Our second key finding is that increased intelligence and morality influenced both perceived trustworthiness and dangerousness of AI, but that these patterns were not uniform. For perceived trustworthiness, we found that across all studies, higher intelligence not only consistently increased expectations of greater morality but also enhanced perceived trustworthiness. Yet we also find that while changes in both intelligence and morality led to increased trustworthiness, it was increased morality that had a stronger effect on enhancing perceived trustworthiness. Moreover, we found that increased intelligence and morality did not impact perceptions of trustworthiness in humans and AI equally. Study 2b showed that increased intelligence had a much stronger effect on trustworthiness for AI than it did for humans. Study 3 found that perceived moral motivation was a stronger predictor of perceived trustworthiness in both humans and AI, but that moral competence was particularly important for AI compared to humans, reflecting the view that intelligent machines are primarily judged by their ability to understand and apply moral reasoning. These findings cohere with a growing consensus that trust in AI is driven by both perceptions of AI performance (intelligence) and morality (Claessens & Everett, 2026; Lee & See, 2004; Malle & Ullman, 2021; see Everett et al. 2026 for a review). Our results cohere with these views while also highlighting the particular importance of morality, potentially reflecting the growing socially embeddedness and consequential nature of AI in everyday life. Most importantly, our results highlight that performance and morality cues are not distinct in perceptions, even if they are in practice. To understand trust in AI, our results highlight how we need to consider not only how perceived trustworthiness depends on perceptions of intelligence and morality, but how these perceptions are not independent themselves and can in fact – mistakenly – be reinforced.

While we found that this perceived "moral illusion" of intelligence led to greater trust, we also found distinct – but more changeable - patterns for perceptions of AI safety, or how dangerous AI would be. While we observed in some studies that manipulations of intelligence led to increased perceptions of danger (Pilot Studies 1a, 2a and 2b, and Study 3), in other studies in reduced perceived danger (Pilot Study B, Study 2b), and in other studies it had no significant effect on perceived danger from AI (Study 1a, Study 2a). Moreover, these results were again more nuanced than what would be expected if our results could be explained simply through a basic halo effect, because across studies, higher morality tended to reduce perceptions of danger for both humans and AI, but higher intelligence reduced fear only of humans, not AI. This suggests that while increased intelligence might lead to perceptions of trustworthiness, it does not necessarily lessen concerns about the potential harm or risk associated with hypothetical future super intelligent agents.

This pattern can be understood through the lens of established models of trust. Mayer et al. (1995) distinguished three antecedents of trust: ability, benevolence, and integrity. Our decomposition of morality into competence and motivation maps onto this framework, with moral competence corresponding broadly to the ability to navigate moral situations and moral motivation corresponding to benevolence and integrity. Viewed this way, our finding that perceived moral motivation was a stronger predictor of trust than moral competence (Study 3) suggests that, for AI, perceived benevolence may matter more for trust than perceived ability. This finding has implications for how AI developers communicate about their systems. Moreover, our central finding, that intelligence drives attributions of both moral competence and moral motivation, suggests that these supposedly distinct antecedents of trust may not be perceived as independent by laypeople, at least in the context of AI.

Our mixed findings on perceived danger complement prior work showing that capable robots and AI systems are perceived as threatening to human resources, safety, and identity (Yogeeswaran et al., 2016; Złotowski et al., 2017). In that literature, greater competence straightforwardly increases perceived threat. Our results suggest a more complex picture: while intelligence may increase perceived danger through increased capability, it simultaneously increases perceived moral motivation, which reduces perceived danger. These opposing forces may explain the inconsistent effects of intelligence on danger across our studies, and suggest that the intelligence-morality inference could partially buffer the threat response that competent AI would otherwise elicit potentially leading to an underestimation of AI risk.

Some approaches to trust in AI tend to implicitly treat perceived danger as the inverse of trust. For example, the widely used Trust in Automation Scale (Jian et al., 2000) includes positively-scored items like "I can trust the system" along with reverse-coded items like "The system's action will have a harmful or injurious outcome." Our results, however, suggest that trust and danger are distinct, or at least can be the result of distinct dual processes: intelligence may increase competence, which can lead to more trust; but also increases the capacity for harm, leading to more danger. As well as highlighting the importance of treating trust and danger as distinct-but-related consequences, our work highlights a potential concern with perceptions of AI dangerousness. For humans, perceptions of morality had a more pronounced effect on trustworthiness than for AI, where low-morality humans were viewed as especially dangerous, potentially due to an intuitive attribution of malicious intent or cruelty to such individuals. In contrast, low-morality AI may be perceived as lacking moral competence—an inability to discern or understand moral principles—rather than as harbouring intentional malevolence.

If people mainly appear to perceive danger when an agent desires to harm them explicitly and people expect higher intelligence to reduce motivation to harm, this highlights a clear concern in line with the Orthogonality Thesis. Highly competent AI can be dangerous even without any "bad" motivations, because it is potentially harder to notice if an agent has mundane or alien goals, such as the inscrutable goals of an agent driven by deep neural networks. By increasing our understanding of how people perceive the relationship between intelligence and morality in humans, then, these findings have concerning implications for AI safety, and in particular the epistemic risk of uncritical narratives around the technological progress of AI.

6.4 Limitations and Future Directions

These experiments present clear evidence that there is a perceived link between intelligence and morality in our sample and they provide novel insights into how people conceptualize the relationship between intelligence, morality, and trustworthiness in AI and human agents. However, as with any research, there are limitations to note.

Although scenarios involving augmented AI intelligence or super-intelligent agents were designed to be relatable, they inevitably simplify the complexity of real-world AI systems and may not fully capture how people think about intelligence and morality in real-world interactive

AI advancements. Our vignettes described AI systems in general terms and did not specify particular architectures or developmental pathways. Future work could examine whether the intelligence-morality inference is moderated by the type of AI system for instance, whether people draw different inferences about systems based on large language models, robotics, or hypothetical whole brain emulation. Relatedly, our studies rely on self-report – as is common in social cognition literature but it would be interesting for future work to explore behavioural indications of trust from descriptions of morality and intelligence, such as trust-based interactions with AI systems in simulated environments. Future work could extend these findings by manipulating perceived intelligence and morality through direct behavioural interactions rather than vignette descriptions, for example by having participants observe AI performance in trust games (cf. Purcell et al., 2025), moral decision tasks (Köbis et al., 2025), or advice-giving contexts (Landes et al., 2026), and examining whether behavioural demonstrations of intelligence produce the same moral inferences as described intelligence. It would be interesting for future studies to explore in more detail the mechanisms by which people learn to assume a connection between intelligence and morality, for humans and AI alike.

While our results suggest that a simple halo effect cannot fully explain these mechanisms for either humans or AI, there may be other explanations that could be explored in future research. One possibility is that participants are exhibiting an anthropomorphic bias based on the idea that human morality and intelligence do go together, and they therefore simply apply this to AI. While this is difficult to explain with the different effects for humans and AI, it would be interesting to explore how people think about the relationship between intelligence and morality in other kinds of non-human agents like animals or even hypothetical aliens. Another possibility is that our results emerge from teleological assumptions that AI is designed for a purpose and so it's (assumed to be ethical) developers would help ensure it acts ethically - though this is perhaps more difficult to explain for our findings on humans (unless participants are implicitly thinking about divine and evolutionary purposes), and it is also more difficult to explain given the augmentation paradigm in which the source of the change was identified as only a specific intelligence-based machine learning breakthrough, and not more general developer-led values.

It is possible that participants anchor scale midpoints differently for human and AI targets on general rating scales (e.g., interpreting the midpoint as reflecting an 'average human' vs. an 'average AI'), which could complicate direct comparisons of absolute rating levels across agent types. Importantly, our central claims do not depend on equivalent anchoring. We show convergent results across between-subjects paradigms (where participants read about a human or AI described as high or low in intelligence), within-subjects augmentation paradigms (where participants rate a system's morality before and after an intelligence augmentation), and explicit change paradigms (where participants directly rate how much morality would change as a result of increased intelligence). These designs test the within-agent inference from intelligence to morality, which does not require that participants anchor the scales equivalently for humans and AI. While some cross-agent comparisons of effect magnitude could in principle be influenced by differential anchoring, the core finding, that intelligence increases perceived morality within each agent type, is robust to this concern. Moreover, across all studies we collected secondary measures in which participants rated the agent's traits relative to an 'average person' baseline (e.g., 'Compared to an average person, how moral do you think [OmegaAI/Tom] is?'), providing a common anchor for both agent types. These measures, reported in the Supplementary Material, produced qualitatively consistent results, suggesting that differential anchoring does not substantially alter the pattern of findings.

Previous work looking at views related to artificially enhanced humans notably have findings pointing in the opposite direction from the pattern we observe: superhuman levels of cognitive ability achieved through neurotechnological enhancement were associated with perceived immorality and dehumanization (Koverola et al., 2022). One possible explanation is that human

enhancement triggered purity and naturalness concerns in those studies that are not applicable to AI systems. Future work would benefit from focusing further on this divergence.

Our studies manipulated intelligence and morality separately rather than simultaneously, but it would be interesting for future work to consider a fully crossed design, allowing for a more precise estimate of the relative contributions of each dimension to perceptions of the other, clarifying whether intelligence is perceived as necessary but not sufficient for morality.

Finally, our sample mainly consisted of British participants. Cross-cultural studies could investigate whether our orthogonality-resistant intuitions about intelligence and morality are a universal phenomenon or whether it is shaped by differing cultural attitudes toward morality, intelligence, and technology.

6.5 Implications

Together, our results provide compelling evidence for lay intuitions about intelligence and morality being connected in artificial intelligence, suggesting that people intuitively associate greater intelligence with not only the cognitive abilities that are deeply intertwined with moral reasoning but also the purely motivational components of morality. This cannot be explained by the possibility that people expect increases in intelligence to merely coincide with increases in morality, since it occurs in both within-subjects augmentation paradigms (Pilot Studies 2a and 2b, and Studies 1b and 2b) and in explicit questions about how this change would directly affect another trait (Study 3). It cannot be explained by the possibility that people are merely thinking of an AI's morality in terms of competences like predicting or repeating back moral norms, because people also perceive an AI that becomes more able to achieve its own goals would also *care* more about humans too by an increase in its moral motivation (Studies 2a, 2b, and 3).

These findings have important implications for how AI advancements are perceived in the public sphere. Media narratives often emphasize rapid progress in AI intelligence. If people are resistant to orthogonality, then this may implicitly suggest that such progress will inherently and automatically lead to safer or more ethical systems. If people are particularly impressed with their interactions with large language models like OpenAI's ChatGPT, Google's Gemini, or Anthropic's Claude, and commentators frequently highlight the impressive leaps in these systems' abilities to engage in what appear to be complex reasoning, this may create an undue sense of optimism about their moral reliability either now or in the future. For example, stories about AI being used to resolve ethical dilemmas—such as applications in healthcare to allocate scarce medical resources (Drezga-Kleiminger et al., 2023; Vinay et al., 2021)—often highlight the AI's capacity to analyze data and make "rational" decisions but rarely address whether such systems genuinely understand moral norms or adhere to them in ways aligned with human values (Bélisle-Pipon et al., 2022) much less how to think about the goals of complex systems based on deep neural networks, which may forever escape full human understanding (Castelvecchi, 2016; Eschenbach, 2021; Myers & Chater, 2024). Similarly, discussions around autonomous vehicles often underscore their advanced decision-making capabilities (Bonnefon et al., 2016; Takaguchi et al., 2022), potentially leading to people assuming that improved intelligence translates to safer, more ethical driving decisions without fully accounting for the moral trade-offs these vehicles might face in real-world scenarios (Awad et al., 2018).

If people believe that simply by making AI smarter, it automatically becomes safer, they may be less inclined to demand, and policymakers less inclined to require, the continued investment in safety engineering that actually produces the alignment they observe. In this way, the intelligence-morality inference could become a self-undermining prophecy as it is approximately accurate today precisely because of safety work that it may discourage in the future.

Optimistic media framing, even if it truly does reflect well how advanced current systems are becoming or how advanced they may become in the future, may inadvertently reinforce the very bias our experiments reveal: the intuitive but potentially unwarranted assumption that advancements in intelligence inherently come with corresponding advancements in morality. By failing to scrutinize whether and how AI systems are explicitly designed to address moral concerns, such narratives may lead to complacency about AI safety. As our findings suggest, people's natural resistance to orthogonality might cloud their ability to critically assess whether AI systems are truly aligned with human values, leading to misplaced trust in these technologies.

6.6 Conclusions

Is an intelligent machine guaranteed to be a moral machine? There are good reasons to assume not. But do ordinary people perceive there to be a necessary connection between intelligence and morality in artificial agents, and what are the consequences of this on trust and danger? In this paper we studied the lay intuitions about intelligence and morality in AI, showing across a series of pre-registered studies that lay perceptions of intelligence and morality are tightly intertwined, with important consequences for how people think about AI development and safety. Across multiple experiments, we consistently found that higher intelligence in AI was associated with greater perceptions of morality and trustworthiness, in contrast to the philosophical and theoretical independence of these traits as outlined by the Orthogonality Thesis. These findings highlight the risks of optimistic narratives that oversell (or even accurately sell) the intelligence of AI and how they might lead to misplaced trust in AI systems. As AI continues to advance and narratives around its capabilities proliferate, it is critical to address misconceptions so as to foster more realistic and informed public discourse about the risks and benefits of AI technologies. Our findings emphasize the importance of incorporating a balanced understanding of intelligence and morality in AI design and policymaking, ensuring that public trust is based on actual safety measures rather than optimistic or anthropocentric biases.

7. Author Contributions (CRediT)

SM: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing.

JACE: Formal Analysis, Funding Acquisition, Methodology, Supervision, Validation, Writing – Review & Editing, Visualization.

8. Acknowledgments

This work was generously supported by funding from the Economic and Social Research Council (ES/V015176/1) and a Philip Leverhulme Prize (PLP-2021-095) awarded to JACE.

9. References

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751–763. https://doi.org/10.1037/0022-3514.93.5.751

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition. In *Advances in Experimental Social Psychology* (Vol. 50, pp. 195–255). Elsevier. https://doi.org/10.1016/B978-0-12-800284-1.00004-7

Adams, E. M. (1980). Gewirth on Reason and Morality. *The Review of Metaphysics*, *33*(3), 579–592.

Aharoni, E., Abdulla, S., Allen, C. H., & Nadelhoffer, T. (2022). Ethical implications of neurobiologically informed risk assessment for criminal justice decisions: A case for pragmatism. In F. De Brigard & W. Sinnott-Armstrong (Eds.), *Neuroscience and Philosophy* (pp. 161–194). The MIT Press. https://doi.org/10.7551/mitpress/12611.003.0010

Anthropic. (2025). In *System cards*. https://www.anthropic.com/system-cards

Armstrong, S. (2013). General purpose intelligence: Arguing the orthogonality thesis. *Analysis and Metaphysics*, (12), 68–84.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Barber, B. (1983). *The Logic and limits of trust*. Rutgers University Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bélisle-Pipon, J.-C., Monteferrante, E., Roy, M.-C., & Couture, V. (2022). Artificial intelligence ethics has a black box problem. *AI & Society, 38*, 1507–1522. https://doi.org/10.1007/s00146-021-01380-0

Black, M. (1964). The gap between "is" and "should." *The Philosophical Review*, *73*(2), 165–181. https://doi.org/10.2307/2183334

Blackburn, S. (1998). *Ruling passions: A theory of practical reasoning*. Oxford University Press.

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

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*(1), 1–31.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2), 71–85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. In *Advances in Experimental Social Psychology* (Vol. 64, pp. 187–262). Elsevier. https://doi.org/10.1016/bs.aesp.2021.03.001

Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. *ACM Computing Surveys*, *55*(3), 1–38. https://doi.org/10.1145/3502289

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Christian, B. (2021). *The alignment problem: How can artificial intelligence learn human values?* Atlantic Books.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *NIPS'17*, 4302–4310.

Claessens, S., & Everett, J. A. C. (2026). *Trust in artificial intelligence is agent-specific and multidimensional*. https://osf.io/preprints/psyarxiv/4y6xw_v1

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

De Neys, W., & Raoelison, M. (2025). Humans and LLMs rate deliberation as superior to intuition on complex reasoning tasks. *Communications Psychology*, *3*(1), 141. https://doi.org/10.1038/s44271-025-00320-8

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1), 28083. https://doi.org/10.1038/s41598-024-79531-8

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations*, *13*(2), 123–139. https://doi.org/10.1177/001872676001300202

Drezga-Kleiminger, M., Demaree-Cotton, J., Koplin, J., Savulescu, J., & Wilkinson, D. (2023). Should AI allocate livers for transplant? Public attitudes and ethical considerations. *BMC Medical Ethics*, *24*(1), 102. https://doi.org/10.1186/s12910-023-00983-0

Dubois, N., & Beauvois, J.-L. (2005). Normativeness and individualism. *European Journal of Social Psychology*, *35*(1), 123–146. https://doi.org/10.1002/ejsp.236

Duff, A. (1977). Psychopathy and Moral Understanding. *American Philosophical Quarterly*, *14*(3), 189–200.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Esmaeilzadeh, P. (2020). Use of AI-based tools for healthcare purposes: A survey study from consumers' perspectives. *BMC Medical Informatics and Decision Making*, *20*(1), 170. https://doi.org/10.1186/s12911-020-01191-1

Everett, J. A. C., Claessens, S., Knöchel, T.-D., & Reinecke, M. G. (2026). Principles for understanding trust in artificial intelligence. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-026-00562-1

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gewirth, A. (1988). The justification of morality. *Philosophical Studies*, *53*(2), 245–262. https://doi.org/10.1007/BF00354643

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy & Technology*, *31*(2), 169–188. https://doi.org/10.1007/s13347-017-0285-z

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Grinschgl, S., Berdnik, A.-L., Stehling, E., Hofer, G., & Neubauer, A. C. (2023). Who wants to enhance their cognitive abilities? Potential predictors of the acceptance of cognitive enhancement. *Journal of Intelligence*, *11*(6), 109. https://doi.org/10.3390/jintelligence11060109

Grinschgl, S., Tawakol, Z., & Neubauer, A. C. (2022). Human enhancement and personality: A new approach towards investigating their relationship. *Heliyon*, *8*(5). https://doi.org/10.1016/j.heliyon.2022.e09359

Herpertz, S. C., & Sass, H. (2000). Emotional deficiency and psychopathy. *Behavioral Sciences & the Law*, *18*(5), 567–580. https://doi.org/10.1002/1099-0798(200010)18:5<567::AID-BSL410>3.0.CO;2-8

Hume, D. (1896). *A treatise of human nature*. Clarendon Press.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kannegieter, T. (2024). AI timelines and national security: The obstacles to AGI by 2027. *Lawfare*. https://www.lawfaremedia.org/article/ai-timelines-and-national-security--the-obstacles-to-agi-by-2027

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology*, *89*(1), 104–118. https://doi.org/10.1037/0021-9010.89.1.104

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, *646*(8083), 126–134. https://doi.org/10.1038/s41586-025-09505-x

Koverola, M., Kunnari, A., Drosinou, M., Palomäki, J., Hannikainen, I. R., Jirout Košová, M., Kopecký, R., Sundvall, J., & Laakasuo, M. (2022). Treatments approved, boosts eschewed: Moral limits of neurotechnological enhancement. *Journal of Experimental Social Psychology*, *102*, 104351. https://doi.org/10.1016/j.jesp.2022.104351

Laakasuo, M., Kunnari, A., Francis, K., Košová, M. J., Kopecký, R., Buttazzoni, P., Koverola, M., Palomäki, J., Drosinou, M., & Hannikainen, I. (2025). Moral psychological exploration of the asymmetry effect in AI-assisted euthanasia decisions. *Cognition*, *262*, 106177. https://doi.org/10.1016/j.cognition.2025.106177

Laakasuo, M., Palomäki, J., Kunnari, A., Rauhala, S., Drosinou, M., Halonen, J., Lehtonen, N., Koverola, M., Repo, M., Sundvall, J., Visala, A., & Francis, K. B. (2023). Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology*, *53*(1), 108–128. https://doi.org/10.1002/ejsp.2890

Landes, E., Francis, K. B., & Everett, J. A. C. (2026). People defer to AI moral advice, but not blindly. *Cognition*, *272*, 106504. https://doi.org/10.1016/j.cognition.2026.106504

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Maibom, H. L. (2008). The mad, the bad, and the psychopath. *Neuroethics*, *1*(3), 167–184. https://doi.org/10.1007/s12152-008-9013-9

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124. https://doi.org/10.1145/2696454.2696458

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction* (pp. 3–25). Elsevier. https://doi.org/10.1016/B978-0-12-819472-0.00001-0

Marcus, G. F. (2018). *Deep learning: A critical appraisal*. arXiv. https://doi.org/10.48550/arXiv.1801.00631

Marcus, G. F. (2024). *Taming silicon valley: How we can ensure that AI works for us*. MIT Press.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

Maynard, M. (2024). Marketers, let's stop overhyping what AI can do. *Forbes Agency Council*. https://www.forbes.com/councils/forbesagencycouncil/2024/09/25/marketers-lets-stop-overhyping-what-ai-can-do/

McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *iScience*, *26*(8), 107256. https://doi.org/10.1016/j.isci.2023.107256

Myers, S., & Chater, N. (2024). *Interactive explainability: Black boxes, mutual understanding and what it would really mean for AI systems to be as explainable as people*. https://doi.org/10.31234/osf.io/ha37x

Nagel, T. (2016). *The possibility of altruism*. Princeton University Press. https://doi.org/10.2307/j.ctt1ggjkt5

Narayanan, A., & Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Conference Companion on Human Factors in Computing Systems - CHI '94*, 204. https://doi.org/10.1145/259963.260288

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250–256. https://doi.org/10.1037/0022-3514.35.4.250

Omohundro, S. M. (2018). The basic AI drives. In *Artificial Intelligence Safety and Security* (pp. 47–55). Chapman; Hall/CRC.

OpenAI. (2023). *Planning for AGI and beyond*. https://openai.com/index/planning-for-agi-and-beyond/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pangle, L. S. (2014). *Virtue is knowledge: The moral foundations of Socratic political philosophy*. University of Chicago Press.

Plato. (1997). Protagoras (S. Lombardo & K. Bell, Trans.). In J. M. Cooper & D. S. Hutchinson (Eds.), *Plato: Complete works* (pp. 746–790). Hackett Publishing.

Purcell, Z. A., & Bonnefon, J.-F. (2023). Research on artificial intelligence is reshaping our definition of morality. *Psychological Inquiry*, *34*(2), 100–101. https://doi.org/10.1080/1047840X.2023.2248857

Purcell, Z. A., Jakesch, M., Dong, M., Nussberger, A.-M., & Köbis, N. (2025). Writing with AI boosts trust-building efficiency. *iScience*, *28*(12), 114092. https://doi.org/10.1016/j.isci.2025.114092

Putnam, H. (2004). *The collapse of the fact/value dichotomy: And other essays*. Harvard University Press.

Railton, P. (1986). Moral realism. *The Philosophical Review*, *95*(2), 163–207.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Reuters. (2024). *Tesla's Musk predicts AI will be smarter than the smartest human next year.* https://www.reuters.com/technology/teslas-musk-predicts-ai-will-be-smarter-than-smartest-human-next-year-2024-04-08/

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283–294. https://doi.org/10.1037/h0026086

Roser, M. (2023). AI timelines: What do experts in artificial intelligence expect for the future? In *Our World in Data*. https://archive.ourworldindata.org/20251125-173858/ai-timelines.html

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Prentice Hall.

Santas, G. (1971). Socrates at work on virtue and knowledge in Plato's *Laches*. In G. Vlastos (Ed.), The Philosophy of SOCRATES. Modern Studies in Philosophy (pp. 177–208). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-86199-6_9

Shafer-Landau, R. (1998). Moral motivation and moral judgment. *The Philosophical Quarterly*, *48*(192), 353–358.

Shafer-Landau, R. (2003). *Moral realism: A defence*. Oxford University Press.

Smith, M. (1994). *The moral problem*. Blackwell.

Stellar, J. E., & Willer, R. (2018). Unethical and inept? The influence of moral information on perceptions of competence. *Journal of Personality and Social Psychology*, *114*(2), 195–210. https://doi.org/10.1037/pspa0000097

Sundvall, J., Drosinou, M., Hannikainen, I., Elovaara, K., Halonen, J., Herzon, V., Kopecký, R., Jirout Košová, M., Koverola, M., Kunnari, A., Perander, S., Saikkonen, T., Palomäki, J., & Laakasuo, M. (2023). Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas. *European Journal of Social Psychology*, *53*(4), 779–804. https://doi.org/10.1002/ejsp.2936

Surdel, N., Bigman, Y. E., Shen, X., Lee, W.-Y., Jung, M. F., & Ferguson, M. J. (2024). Judging robot ability: How people form implicit and explicit impressions of robot competence. *Journal of Experimental Psychology: General*, *153*(5), 1309–1335. https://doi.org/10.1037/xge0001548

Takaguchi, K., Kappes, A., Yearsley, J. M., Sawai, T., Wilkinson, D. J. C., & Savulescu, J. (2022). Personal ethical settings for driverless cars and the utility paradox: An ethical analysis of public attitudes in UK and Japan. *PLOS ONE*, *17*(11), e0275812. https://doi.org/10.1371/journal.pone.0275812

Takemoto, K. (2026). *Scaling laws for moral machine judgment in large language models*. arXiv. https://doi.org/10.48550/arXiv.2601.17637

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Turchin, A. (2019). *AI alignment problem: "Human values" don't actually exist*. https://www.lesswrong.com/posts/ngqvnWGsvTEiTASih/ai-alignment-problem-human-values-don-t-actually-exist

Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & Society*, *35*(1), 147–163. https://doi.org/10.1007/s00146-018-0845-5

Vinay, R., Baumann, H., & Biller-Andorno, N. (2021). Ethics of ICU triage during COVID-19. *British Medical Bulletin*, *138*(1), 5–15. https://doi.org/10.1093/bmb/ldab009

Waser, M. R. (2008). Discovering the foundations of a universal system of ethics as a road to safe artificial intelligence. *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures*, 195–200.

Wiggins, D. (1998). *Needs, values, truth: Essays in the philosophy of value*. Oxford University Press. https://doi.org/10.1093/oso/9780198237198.001.0001

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, *67*(2), 222–232. https://doi.org/10.1037/0022-3514.67.2.222

Wojciszke, B. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology*, *16*(1), 155–188. https://doi.org/10.1080/10463280500229619

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human-Robot Interaction*, *5*(2), 29–47. https://doi.org/10.5898/JHRI.5.2.Yogeeswaran

Yudkowsky, E. (2016). *The AI alignment problem: Why it's hard, and where to start*. https://intelligence.org/files/AlignmentHardStart.pdf

Yudkowsky, E., & Soares, N. (2025). *If anyone builds it, everyone dies: The case against superintelligent AI*. Little, Brown; Company.

Zaim Bin Ahmad, M. S., & Takemoto, K. (2025). Large-scale moral machine experiment on large language models. *PLOS One*, *20*(5), e0322776. https://doi.org/10.1371/journal.pone.0322776

Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies*, *100*, 48–54. https://doi.org/10.1016/j.ijhcs.2016.12.008