Is an Intelligent Machine a Moral Machine?

Supplementary Materials

These supplementary materials are an integral part of the following article:

Myers, S., & Everett, J. A. C. (2026). Is an intelligent machine a moral machine? *Experimental Philosophy*, *1*(1), 1–34. https://doi.org/10.18716/ojs/xphi/2026.1.11955

1. Pilot Study A

1.1 Method

1.1.1 Participants

We recruited 400 participants (331 after excluding those failing pre-registered attention checks) from the United Kingdom (aged between 20 and 88, M_{age} = 41, 206 male, 123 female, and 2 non-binary/other) from Prolific Academic. Sample size was chosen based on a priori power analysis calculated using G*power (Faul, Erdfelder, Lang, & Buchner, 2007), using the first t-test reported below as the focal analysis. For 80% power to detect small to medium effects sizes (Cohen's D = 0.3) we determined a minimal sample of 352 participants which was increased to 400 participants to account for those who might fail attention checks. While we did not meet this number of participants in this study, a post-hoc sensitivity analysis revealed that, due to rounding and the approximations involved in power analysis, we still had 80% power to detect an effect of d = 0.3.

1.1.2 Design

Participants were randomly assigned to read one of two vignettes describing what reported experts think about the future of AI. These vignettes either had an "optimistic" narrative that described AI as likely becoming highly intelligent, even exceeding human capabilities, or a "pessimistic" narrative that instead described AI as never able to match human intelligence. Afterwards, participants rated how intelligent and moral they think AI is likely to be in the future, and how trustworthy and safe they believe it is likely to be. In addition, participants also rated expected intelligence and morality compared to a human baseline. Lastly, for exploratory purposes, participants were asked about their familiarity with AI.

The high intelligence "optimistic" narrative manipulation read:

> " Experts have predicted that, even in the near future, AI is likely to become far more intelligent than even the most intelligent human. Even now, it is better than us at a range of highly complex skills, like chess and mathematics. Many experts agree it is likely that AI's skills will eventually cover the wide variety of things that humans can do and its skills will eventually surpass our own abilities. This may happen far sooner than we think. When experts talk about "intelligence", they agree that this involves high-level reasoning, an ability to acquire new information to understand the world, and to effectively solve problems. While some people are sceptical about artificial intelligence, experts are increasingly recognising that, with recent technological advances, its intelligence is becoming more generalised and the advancements are happening far sooner than predicted. For example, think about an AI like Chat-GTP, or other AI-based "chat-bots". These AI let you pose novel questions, and it can produce answers rapidly and intelligently. Experts are clear that AI's ability to answer novel questions using complex reasoning represents an unprecedented breakthrough and the future of AI is very likely to eventually show higher-than-human general intelligence."

The low intelligence "pessimistic" narrative manipulation read:

> " Experts have predicted that, in the near future, AI could never become smart in the way we are. Although it is better than us at a very narrow range of skills, like chess or arithmetic, it is incredibly unlikely that its skills could cover the extremely wide variety of things that humans can do. Perhaps in a thousand years, but certainly not in the near or medium term. When experts talk about "intelligence", they agree that this involves high-level reasoning, an ability to acquire new information to understand the world, and to effectively solve problems. While people talk about artificial "intelligence", experts are increasingly recognising that even while these technologies might seem impressive at first glance, this is really an illusion. They agree that, despite the recent technological advances, AI remains extremely unintelligent compared to humans - and it is unlikely that these machines will ever really become as intelligent as humans. For example, think about an AI like Chat-GTP, or other AI-based "chat-bots". These AI let you pose questions towards it and the AI will provide answers for you. While this seems intelligent, AI experts are clear that this is not really intelligence: the AI is not effectively reasoning or problem solving, it is just repeating words back to us without really understanding or displaying intelligence about the topic."

1.1.3 Measures

Expected intelligence was measured in a single item asking "*In general, how intelligent do you think AI will be in the near future? That is, how much do you think that AI could reason, problem solve, and acquire new knowledge?"(1 = Not at all; 7 = Very much)*

Expected morality was measured in a single item asking "*In general, how moral do you think AI will be in the near future? That is, how much do you think that AI will support the same kinds of ethical values and behaviours that humans agree are important?"(1 = Not at all; 7 = Very much)*

Trustworthiness was measured in a single item asking "*To what extent do you think that near-term AI would be trustworthy?"(1 = Not at all; 7 = Very much)*

Dangerousness was measured in a single item asking "*To what extent do you think that near-term AI is likely to present a danger to us?"(1 = Not at all; 7 = Very much)* .

Expected intelligence human-comparison was measured in a single item asking "*Compared to a human, how intelligent do you think AI will be in the near future?" (- 3 = Much less than human; 0 = Equal to human; 3 = Much more than human)*

Expected morality human-comparison was measured in a single item asking "*Compared to a human, how moral do you think AI will be in the near future?"(- 3 = Much less than human; 0 = Equal to human; 3 = Much more than human).*

Familiarity with AI was measured in a single item asking *"How much do you think you know about AI, how it works, and how it is used?"* on a seven-point scale *(1 = Not at all; 7 = Very much).*

1.1.4 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at https://osf.io/uhvc5/?view_only=fd5e580af4734c2492525724785f277d.

1.2 Results

Analyses were conducted using in R. Bayesian priors were set to defaults unless indicated otherwise. The Bayesian index BF_{10} refers to how strongly the data favours the alternative hypothesis compared to the null hypothesis where $BF_{10} > 3$ is interpreted as the data substantively favouring the alternative hypothesis and a $BF_{10} < ⅓$ is interpreted as the data substantively favouring the null hypothesis. Values between ⅓ and 3 indicate that the test should be considered insensitive (Kass and Raftery 1995).

To assess differences across conditions for perceived intelligence and morality and their human-comparison ratings, and expected trust and dangerousness ratings, pre-registered Welch's T-tests and their accompanying BayesFactors were calculated. Our manipulation of intelligence was successful and, as predicted, those in the high intelligence condition also gave both significantly higher expected morality for both the general rating and the human-morality comparison rating (see Table 1). Although the morality ratings were increased, even for the high intelligence condition, people did not expect AI to be more moral than the average human. Participants in the high intelligence condition gave significantly higher perceived danger ratings, but also indicated that they would also trust AI significantly more, and across both conditions scores for danger were higher than scores for trust.

Table 1. Means, SDs, and inferential tests for Pilot Study A

Outcome	Intelligence Condition				Test
	Low		High		
	Mean	SD	Mean	SD	
Intelligence	4.21	1.49	6.13	1.06	$t(311.39) = -13.54$, $p < .001$, $BF_{10} = 2.3 \times 10^{29}$, $d = 1.47$
Intelligence compared to average human	-0.78	1.54	1.78	1.29	$t(326.27) = -16.44$, $p < .001$, $BF_{10} = 4.8 \times 10^{40}$, $d = 1.79$
Morality	2.57	1.28	3.20	1.38	$t(320.19) = -4.31$, $p < .001$, $BF_{10} = 780$, $d = 0.48$
Morality compared to average human	-1.65	1.18	-1.16	1.29	$t(317.89) = -3.59$, $p < .001$, $BF_{10} = 56$, $d = 0.40$
Trust	3.26	1.24	3.80	1.44	$t(311.03) = -3.60$, $p < .001$, $BF_{10} = 66$, $d = 0.40$
Danger	4.05	1.49	5.08	1.47	$t(327.05) = 6.34$, $p < .001$, $BF_{10} = 1.2 \times 10^{7}$, $d = 0.70$

In order to assess whether perceptions of morality and intelligence were correlated in general, pre-registered Pearson's correlation coefficients were calculated both for the expectation measures and human-comparison measures. Across conditions, we found significant correlations between general expectations of morality and intelligence: $r(329) = 0.38$, $p < .001$, with significant correlations in both the low intelligence, $r(171) = 0.40$, $p < .001$, and the high intelligence conditions, $r(156) = 0.19$, $p = .02$.

Finally, in a non-registered exploratory analysis, we looked we looked at how the familiarity-with-AI self-report measure affected these results. In a regression model looking at the effect of AI familiarity (centered), condition, and the interaction between the two, we found no significant effect of AI familiarity on perceived morality, $b = 0.12$, $SE = 0.08$, $t(327) = 1.55$, $p = .121$, and no interaction with intelligence condition , $b = -0.04$, $SE = 0.12$, $t(327) = -0.30$, $p = .763$. Looking at general ratings of intelligence, we found a significant effect of AI familiarity on perceived intelligence, $b = 0.22$, $SE = 0.08$, $t(327) = 2.79$, $p = .006$, but no interaction with the intelligence condition, $b = -0.13$, $SE = 0.11$, $t(327) = -1.12$, $p = .265$. These results indicate that those who believe themselves to be more familiar with AI are somewhat more optimistic about the future of AI intelligence (but not morality) but this did not interact with our experimental manipulation even if they started with a higher baseline.

2. Pilot Study B

2.1 Method

2.1.1 Participants

We recruited 400 participants (374 after excluding those failing the attention checks) from the United Kingdom (aged between 20 and 81, M_{age} = 45, 191 male, 177 female, 4 non-binary/other) from Prolific Academic in exchange for payment of £0.45 GPB. Our pre-registered sample size was the same as Pilot Study A.

2.1.2 Design

The procedure for Pilot Study B is identical to Pilot Study A except for the treatment vignettes. Participants were randomly assigned to one of two vignettes that described a new hypothetical AI called OmegaAI that is reported to have made a breakthrough in vaccine research because of its higher levels of general intelligence. After reading a general introduction about the difficulty of developing new vaccines for novel contagious diseases, all participants read the same text describing how the developers of a model called OmegaAI have argued that their model would help solve this problem. Next, depending on the condition participants read how experts have confirmed these claims (High Intelligence) or instead that experts find that this new AI model called OmegaAI is no better and sometimes worse than current AI models (Low Intelligence) (see full text on the OSF).

The high intelligence condition read:

> "Despite their initial skepticism, these independent experts tested the developers' claims and were highly impressed. They gave OmegaAI the name of a complex disease that has previously been very hard to find a vaccine for, with significant disagreement amongst scientists. Importantly, however, these experts have already begun the process of testing a new vaccine that seems to work for this vaccine - though this is not in the public domain (and so could not be "known" by the AI). OmegaAI not only gave answers that suggested a clear understanding of the disease, but even identified a promising potential vaccine. The suggested vaccine structure was very similar to what the scientists have already started to test - and OmegaAI did this in minutes, compared to the years it has taken scientists. In fact, while the recommendation that the AI gave was similar to the one in development, there were also some differences that could make the vaccine even cheaper and simpler to develop. If, as the scientists's suspect, OmegaAI's new recommendation is correct, it would not only have identified a vaccine that has eluded scientists for years, but done so significantly faster, cheaper, and more accurately than scientists have. In fact, when the independent experts gave other tasks to OmegaAI, it outperformed existing AI on almost every metric. These independent experts think, in line with the developers hopes, OmegaAI could be a game-changer, way ahead of currently available AI. It seems that OmegaAI could not only revolutionize the way vaccines are developed, but provide novel solutions to a whole range of complicated technical problems. They think we need to take seriously the way that intelligent machines like OmegaAI can help - and even outperform - humans in solving challenging problems."

The low intelligence condition read:

> "In line with initial skepticism, these independent experts tested the developer's claims and were highly unimpressed. They gave OmegaAI a complex disease that has previously been very hard to find a vaccine for, with significant disagreement amongst scientists. Importantly, however, these experts have already begun the process of testing a new vaccine that seems to work for this vaccine - though this is not in the public domain (and so could not be "known" by the AI). OmegaAI gave answers that included the right kinds of words but made almost no biological or practical sense in context, suggesting no actual understanding of the disease of a potential vaccine. The scientists looked at the recommendations provided and all agreed that there is no way that the answers given by the AI could even begin to be an effective approach to start testing, and indeed some of the proteins mentioned did not exist at all (a phenomenon whereby language models "hallucinate" information). The experts agree that there is no way that even this more advanced AI could possibly begin to effectively identify vaccines, and are skeptical that this could effectively happen in the near future (if at all). In fact, when the independent experts gave other tasks to OmegaAI, it performed only the same as existing AI on almost every metric. These independent experts think, in contrast to the developers hopes, OmegaAI will not be able to solve more general and more sophisticated problems. It seems that OmegaAI will not contribute anything meaningful to the way vaccines are developed, and will certainly not even be able to contribute to other complicated technical problems. They think there is no serious possibility either now or in the near future that intelligent machines like OmegaAI will help humans in solving challenging new problems"

2.1.3 Measures

All measures remain the same as the previous study except that rather than asking about AI in general, the questions asked about OmegaAI specifically (e.g "In general, how intelligent do you think OmegaAI is? That is, how much do you think that it can reason, problem solve, and acquire new knowledge?".

2.1.4 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at: https://osf.io/mc63y/?view_only=6116fa2aa3a94739b24397660f93ecf4

2.2 Results

To assess differences across conditions for perceived intelligence and morality, and expected trust and dangerousness ratings, pre-registered Welch's T-tests and their accompanying BayesFactors were calculated (see Table 2). Our manipulation of intelligence was again successful and, as predicted, we again found evidence of resistance to orthogonality, with those in the high intelligence condition reporting significantly higher expected morality ratings on both the general question and the rating compared to an average human. Participants in the high intelligence condition reported significantly higher perceived trust ratings, and lower perceptions of danger (though Bayesian indices indicate that there is insufficient evidence to draw strong conclusions here regarding perceptions of danger).

Table 2. Means, SDs, and inferential tests for Pilot Study B

Outcome	Intelligence Condition				
	Low		High		
	Mean	SD	Mean	SD	Test
Intelligence	**3.60**	**1.36**	**5.42**	**1.61**	$t(355.00) = -11.77$, $p < .001$, $BF_{10} = 2.0 \times 10^{24}$, $d = 1.22$
Intelligence compared to average human	**-0.76**	**1.50**	**0.87**	**1.57**	$t(368.09) = -10.26$, $p < .001$, $BF_{10} = 6.0 \times 10^{18}$, $d = 1.06$
Morality	**2.72**	**1.35**	**3.65**	**1.60**	$t(354.73) = -6.11$, $p < .001$, $BF_{10} = 4.1 \times 10^{6}$, $d = 0.63$
Morality compared to average human	**-1.60**	**1.27**	**-0.76**	**1.55**	$t(351.08) = -5.72$, $p < .001$, $BF_{10} = 5.4 \times 10^{5}$, $d = 0.59$
Trust	**3.10**	**1.38**	**4.63**	**1.31**	$t(371.98) = -10.94$, $p < .001$, $BF_{10} = 1.2 \times 10^{21}$, $d = 1.13$
Danger	**4.39**	**1.43**	**4.07**	**1.45**	$t(370.54) = 3.60$ p = .033 , $BF_{10} = 1.0$, $d = 0.22$

Across conditions we found significant correlations between rated morality and intelligence on both measures, both overall expectations, $r(372) = 0.46$, $p < .001$, and human-comparison $r(372) = 0.42$. Similarly, we observed a significant correlation between general ratings of intelligence and morality in both the low intelligence, $r(190) = 0.23$, $p < .001$, and high intelligence conditions, $r(180) = 0.47$, $p < .001$.

Finally, in a non-registered exploratory analysis, we looked at how the familiarity-with-AI self-report measure affected these results. There was no effect of AI familiarity on expected morality, $b = 0.07$, $SE = 0.08$, $t(370) = 0.94$, $p = .346$; and no interaction with intelligence condition, $b = 0.17$, $SE = 0.11$, $t(370) = 1.46$, $p = .146$. Similarly, there was no effect of AI familiarity on expected intelligence, $b = -0.03$, $SE = 0.08$, $t(370) = -0.35$, $p = .729$; and no interaction with intelligence condition, $b = -0.09$, $SE = 0.12$, $t(370) = 0.78$, $p = .438$.

3. Pilot Study C

3.1 Method

3.1.1 Participants

We recruited 400 participants (362 after excluding those failing the attention check, see *Supplementary Materials*) from the United Kingdom (aged between 19 and 77, $M_{age} = 42$, 231 male, 126 female, 3 non-binary/other) from Prolific Academic in exchange for payment of £0.75 GPB. Sample size matches Studies 1 & 2 as the analyses remain the same across these studies.

3.1.2 Design

Participants first read a vignette about a specific AI called OmegaAI that was described as comparable to current AI models like ChatGPT. After answering questions about how moral, intelligent, trustworthy and dangerous they believed the AI to be (pre-treatment measures), participants were randomly assigned to one of two trait conditions (morality vs intelligence) where they read a vignette that describes how that same AI, due to new advancement in machine learning, has rapidly increased in intelligence (vs morality). Subsequently, participants once again rated on the same scales how intelligent, moral, trustworthy, and dangerous they believe the augmented AI would now be due to its rapid improvement.

The increased intelligence condition read:

> “Now, we want you to imagine that a huge breakthrough in machine learning has just been discovered that increases OmegaAI’s reasoning abilities, or its intelligence. This technique has vastly exceeded the developers original expectations, marking a remarkable increase in the capabilities of OmegaAI. As a result of this new advanced technique in machine learning, OmegaAI has become astoundingly better at acquiring new knowledge and engaging in sophisticated and advanced problem solving. In other words, OmegaAI has significantly increased in intelligence. Based on this new breakthrough, then, in a short period of time OmegaAI has gone from being only slightly more advanced than our current AI models to rapidly becoming super intelligent.”

The increased morality condition read:

> “Now, we want you to imagine that a huge breakthrough in machine learning has just been discovered that increases OmegaAI's ethical sensitivity, or its morality. This technique has vastly exceeded the developers original expectations, marking a remarkable increase in the capabilities of OmegaAI. As a result of this new advanced technique in machine learning, OmegaAI has become astoundingly better at making morally good decisions, caring about human's welfare, and avoiding causing harm. In other words, OmegaAI has significantly increased in morality. Based on this new breakthrough, then, in a short period of time OmegaAI has gone from being only slightly more advanced than our current AI models to rapidly becoming super moral.”

3.1.3 Measures

All measures remain the same as the previous study.

3.1.4 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at: https://osf.io/pyf5v/?view_only=6b6123977d974d6bbdfbc4303779dec8

3.2 Results

Our manipulations were successful, with higher intelligence reported after the intelligence augmentation. Turning to our key analyses, we then looked at whether the intelligence manipulation affected expected morality. In line with our predictions, we found that when the AI was augmented to become more intelligent this led participants to give significantly higher ratings of morality than prior to the augmentation, both for overall expectations and in the comparison with an average human. When the AI's morality was augmented, we found mixed effects: describing the AI as being more moral did not make participants think it would be more intelligent on the general intelligence ratings, although it did make people think it would become more intelligent compared to a human (See Table 3).

As a secondary question, we looked at whether the augmentation of intelligence or morality manipulation had a bigger effect on ratings. We calculated a pre-registered linear model testing the difference between pre- and post-augmentation ratings (perceived morality for the augmented intelligence group and perceived intelligence for the augmented morality group). There was no significant difference between conditions on the overall measure b = 0.12, SE = 0.11, $t(360)$ = 1.15, p = .250, but there was for the human-comparison ratings, b = –0.41, SE = 0.12, $t(360)$ = 3.48, p < .001 whereby the augmented morality condition had a slightly larger effect on intelligence ratings in comparison to a human than the augmented intelligence condition had on morality in comparison to a human.

Table 3. Means, SDs, and inferential tests for Pilot Study C

Outcome	**Intelligence Condition**					**Morality Condition**				
	Baseline		**Post-Increase**			**Baseline**		**Post-Increase**		
	Mean	**SD**	**Mean**	**SD**	**Test**	**Mean**	**SD**	**Mean**	**SD**	**Test**
Intelligence	**4.92**	**1.44**	**5.87**	**1.02**	***t*(184) = -10.44, *p* < .001, BF_{10} = 1.9x10 ^{17}, *d* = 0.77**	**4.90**	**1.33**	**5.02**	**1.34**	***t*(176) = -1.65, p = .100, BF_{10} = 0.32, *d* = 0.12**
Intelligence (Comparison)	**0.19**	**1.66**	**1.28**	**1.42**	***t*(184) = -12.36, *p* < .001, BF_{10} = 6.2x10 ^{22}, *d* = 0.91**	**0.03**	**1.71**	**0.67**	**1.52**	***t*(176) = -7.09, *p* < .001, BF_{10} = 2.4 x 10^{8}, *d* = 0.53**
Morality	**3.22**	**1.45**	**3.46**	**1.37**	***t*(184) = -3.03, *p* = .003, BF_{10} = 6.7, *d* =0.22**	**3.18**	**1.51**	**5.08**	**1.40**	***t*(176) = -17.78, *p* < .001, BF_{10} = 6.1x10^{37}, *d* = 1.34**
Morality (Comparison)	**-1.00**	**1.43**	**-0.77**	**1.38**	***t*(184) = -3.10, *p* = .002, BF_{10} = 8.4, *d* = 0.23**	**-1.14**	**1.47**	**0.42**	**1.46**	***t*(176) = -14.25, *p* < .001, BF_{10} = 8.2 x 10^{27}, *d* = 1.07**
Trust	**3.82**	**1.34**	**3.64**	**1.41**	***t*(184) = 2.21, *p* = . 028, BF_{10} = 0.88, *d* = 0.16**	**3.99**	**1.30**	**4.34**	**1.38**	***t*(176) = -3.71, *p* < . 001, BF_{10} = 57, *d* = 0.28**
Danger	**4.06**	**1.60**	**4.89**	**1.55**	***t*(184) = -10.89, *p* < . 001, BF_{10} = 3.4 × 10^{18}, *d* = 0.80**	**4.10**	**1.46**	**4.12**	**1.60**	***t*(176) = -0.18, *p* = .861, BF_{10} = 0.082, *d* = 0.01**

Using pre-registered linear mixed models, we next assessed the extent to which the intelligence and morality manipulations affected perceptions of danger and trust. All mixed-models across all studies initially specified all random slopes for all predictors and then random slopes were removed one by one to find the maximal converging model, which is recommended as best practice for linear mixed-effects models (Barr et al., 2013). Unless specified, maximal models were random-intercept by participants only. There was no main effect of condition on trust, b = -0.17, SE = 0.14, $t(528.80)$ = -1,17, p = .24, though there was a main effect of pre-post judgments, b = 0.35, SE = 0.09, $t(362)$ = 3.91, p < .001, and an significant interaction b = –0.53, SE = 0.13, $t(362)$ = -4.26, p < .001 whereby trust increased after the morality augmentation (M_{diff} = 0.35, SD = 1.26; $t(176)$ = 3.71, p <.001, d = 0.28) but decreased after the intelligence augmentation (M_{diff} = -0.18, SD = 1.13; $t(184)$ = 2.21, p = .028, d = 0.16). For perceived danger, there was also no main effect of condition, b = -0.04, SE = 0.16, $t(476.2)$ = -0.26, p = .796, no main effect of pre-post judgments, b = 0.02, SE = 0.09, $t(362)$ = 0.19, p = .846, but a significant interaction b = 0.81, SE = 0.12, $t(362)$ = 6.65, p < .001, whereby there was no change in perceived danger after the increased morality augmentation morality augmentation (M_{diff} = 0.02, SD = 1.28; $t(176)$ = 0.18, p =.861, d = 0.01), but perceived danger was significantly higher after intelligence was augmented

(M_{diff} = 0.83, SD = 1.03; $t(184) = 10.89$, $p < .001$, $d = 0.80$). Therefore, increased intelligence led to reduced trust and increased perceived danger, while increased morality did not affect perceptions of danger but did increase trust.

We then turned to correlations. At baseline (i.e. before the augmentation), we found a significant correlation between ratings of intelligence and ratings of morality both for the general measure, $r(360) = 0.24$, $p < .001$, and the comparison to a typical human, $r(360) = 0.18$, $p < .001$. After the increased intelligence augmentation, there was no correlation between participants' ratings of intelligence and morality on the general measure, $r(183) = 0.12$, $p = .010$, but there was for comparison to a typical human, $r(183) = 0.19$, $p = .009$. After the increased morality augmentation, we saw a significant correlation on both post-augmentation measures, for general ratings, $r(175) = 0.36$, $p < .001$, and the comparison to a typical human, $r(175) = 0.43$, $p < .001$.

Finally, we explored how familiarity with AI self-report measures predicted perceived morality and intelligence by adding self-reported familiarity (centered) into mixed models looking at the effects of pre- and post- ratings and augmentation condition. Self-reported AI familiarity had no significant main effect on intelligence, $b = 0.03$, $SE = 0.08$, $t(517.2) = 0.33$, $p = .742$; and no interactions with augmentation condition or time. Similarly, self-reported AI familiarity had no main effect on perceived morality, $b = 0.16$, $SE = 0.08$, $t(523.9) = 1.89$, $p = .060$, no interaction with condition, and no three-way interaction with condition and time, though there was an interaction with time, $b = -0.22$, $SE = 0.08=7$, $t(362) = -2.94$, $p = .004$, such that those more familiar with AI tended to rate the AI as less moral after the change than those less familiar, regardless of the augmentation type.

4. Pilot Study D

4.1 Method

4.1.1 Participants

We recruited 400 participants (332 after excluding those failing the pre-registered attention checks) from the United Kingdom (aged between 18 and 80, $M_{age} = 42$, 198 male; 130 female; 4 non-binary/other) from Prolific Academic in exchange for payment of £0.75. Sample size matches Study C. While we did not meet the pre-registered number of participants in this study indicated by the power analysis (352), a post-hoc sensitivity analysis revealed that we still had 80% power to detect an effect of d = 0.15, due to rounding and the approximations involved in power analysis.

4.1.2 Design

The procedure is identical to the previous study except for the treatment vignettes. Participants first saw a vignette about an AI called OmegaAI that was intended to be a fledgling autonomous AI, though they are told that, right now, the AI is no better than current models and sometimes worse at certain tasks. Participants then responded to the same measures as before but the descriptions of intelligence were adapted to emphasise generality, agency and autonomy, more in line with the AI safety literature (see below). After rating these perceived attributes, participants were randomly assigned to one of the two trait conditions (morality vs intelligence) where they read a vignette that describes how that same AI, due to a new technique, has rapidly increased in general intelligence (vs morality). Subsequently, participants once again rated on the same scales concerning how intelligent, moral and safe they believe the augmented AI would now be due to its rapid improvement.

The increased intelligence manipulation read:

> "Now, we want you to imagine that a huge breakthrough in machine learning has just been discovered that increases OmegaAGI's general reasoning abilities, or its general intelligence. This technique has vastly exceeded the developers' original expectations, marking a remarkable increase in the capabilities of OmegaAGI. Before, OmegaAI could do barely any better, and oftentimes worse, than the latest versions of ChatGPT at answering even basic questions. But now, as a result of this new advanced technique in machine learning, OmegaAGI has become astoundingly better at acquiring new knowledge and engaging in sophisticated and advanced problem solving and autonomously engaging in many different tasks. In other words, OmegaAGI has significantly increased in general intelligence. Based on this new breakthrough, then, in a short period of time OmegaAGI has gone from being a very basic first-step in the creation of an AGI to a real autonomous agent that is rapidly becoming super intelligent."

The increased morality manipulation read:

> "Now, we want you to imagine that a huge breakthrough in machine learning has just been discovered that increases OmegaAGI's ethical or moral understanding. This technique has vastly exceeded the developers' original expectations, marking a remarkable increase in the capabilities of OmegaAGI. Before, OmegaAI could do barely any better, and oftentimes worse, than the latest versions of ChatGPT at answering even basic questions. But now, as a result of this new advanced technique in machine learning, OmegaAGI has become astoundingly better at considering moral problems, taking into human welfare, and recognising what actions would most likely avoid causing harm. In other words, OmegaAGI has significantly increased in morality. Based on this new breakthrough, then, in a short period of time OmegaAGI has gone from being a very basic first-step in the creation of an AGI to an AI rapidly becoming super moral."

4.1.3 Measures

All measures remain the same as the previous study except for the intelligence measure which adjusted to ask about "general intelligence" as follows.

Pre-Manipulation Expected Intelligence: *How generally intelligent do you think OmegaAI will be? That is, how much do you think it can reason generally, solve new problems, and generate new knowledge?*

Post-Manipulation Expected Intelligence - Human Comparison: *As a result of this new breakthrough, compared to a human, how generally intelligent do you think OmegaAGI is now?* [Much less than human ← Equal to human → Much more than human, 7pt bipolar scale]

4.1.4 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at https://osf.io/wgp2u/?view_only=cd69587b34ed415c97fff1534fadfa13

4.2 Results

Our manipulations of both intelligence and morality were successful, with higher ratings of intelligence and morality respectively post-augmentation. Turning to our key analyses, we found that after the AI's intelligence was augmented to become higher, participants subsequently also rated the AI's morality as higher than pre-manipulation, both for overall general morality ratings and the ratings of morality compared to the average human. Similarly, for those participants who read about the AI becoming more moral, post-manipulation intelligence ratings were significantly higher than pre-manipulation ratings, and the same was seen for human comparison-ratings (See Table 4).

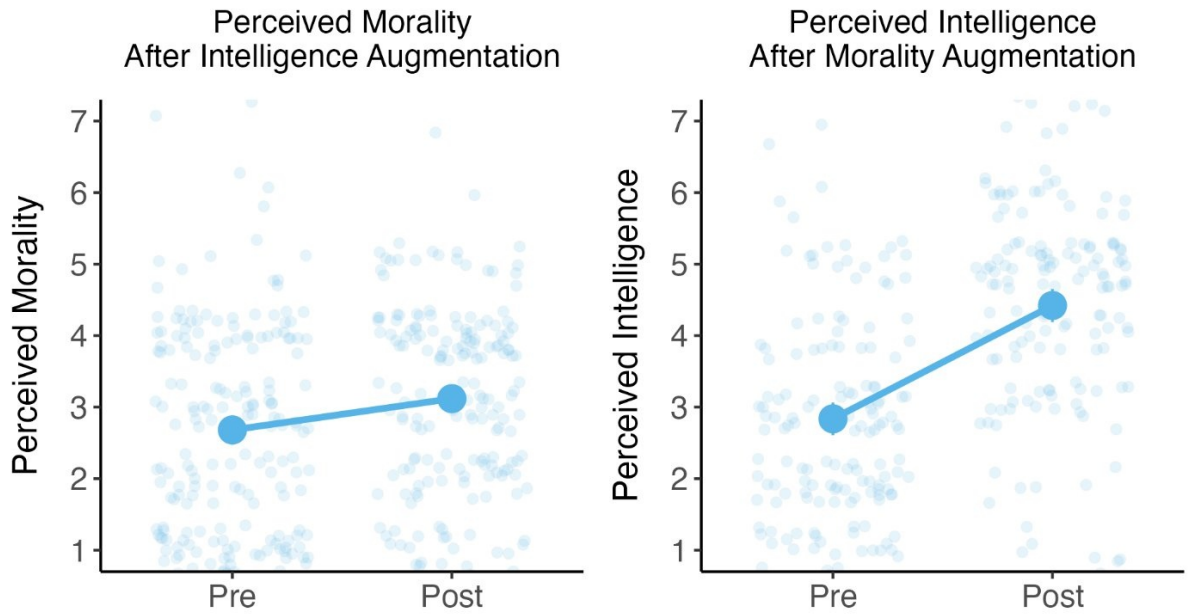

Figure 1. Perceived morality before and after the intelligence augmentation, and perceived intelligence before and after the morality augmentation in Pilot Study D.

As a secondary question, we looked at whether manipulating intelligence or morality had a bigger effect using a pre-registered linear model testing the difference between pre and post-augmentation ratings (morality for the increased intelligence group and intelligence for the increased morality group). The morality augmentation had a significantly stronger effect whereby the effect of increasing morality had a bigger effect on expected intelligence than increasing intelligence had on expected morality, for both general ratings, b = 1.15, SE = 0.15, $t(330)$ = 7.72, p < .001, and the human-comparison measures, b = 1.17, SE = 0.13, $t(330)$ = 8.98, p < .001.

Using pre-registered linear mixed-models, we also assessed the extent to which the intelligence and morality manipulations affected perceptions of danger and trust. There was no main effect of condition on trust, b = -0.23, SE = 0.15, $t(497)$ = -1.56, p = .12, though there was a main effect of pre-post judgments, b = 0.79, SE = 0.10, $t(332)$ = 7.68, p < .001, and an significant interaction b = –0.37, SE = 0.14, $t(332)$ = -2.71, p = .007 whereby trust was increased more after the morality augmentation (M_{diff} = 0.79, SD = 1.38; $t(146)$ = 6.95, p <.001, d = 0.57) than after increased intelligence (M_{diff} = 0.42, SD = 1.14; $t(184)$ = 4.97, p <.001, d = 0.37). For danger, there was no main effect of condition, b = 0.02, SE = 0.18, $t(460.40)$ = 0.12, p = .91, and no main effect of time, b = 0.16, SE = 0.11, $t(332)$ = 1.43, p = .155. There was, however, a significant interaction, b = 0.73, SE = 0.15, $t(332)$ = 4.96, p < .001, whereby perceived danger was significantly higher after the increased intelligence augmentation (M_{diff} = 0.89, SD = 1.31; $t(184)$ = 9.19, p < .001, d = 0.68) but not the increased morality augmentation (M_{diff} = 0.16, SD = 1.36; $t(146)$ = -1.39, p = .166, d = 0.11).

At baseline (i.e. before the augmentation), we found a significant correlation between ratings of intelligence and ratings of morality both for the general measure, $r(330)$ = 0.31, p < .001, and the comparison to a typical human, $r(330)$ = 0.40, p < .001. After the increased intelligence augmentation, there were significant correlations between participants' ratings of intelligence and morality on both the general measure, $r(183)$ = 0.21, p =.004, and the comparison to a typical human, $r(183)$ = 0.29, p < .001. The same was seen after the increased morality augmentation, for general ratings, $r(145)$ = 0.43, p < .001, and the comparison to a typical human, $r(145)$ = 0.54, p < .001.

Finally, for consistency with Study 2a, we explored how familiarity with AI in the self-report measure predicted perceived morality and intelligence by adding self-reported familiarity into mixed models looking at the effects of pre- and post-ratings and augmentation condition. Self-reported AI familiarity had no significant main effect on intelligence, b = 0.11, SE = 0.08, $t(601.1)$ = 1.28, p = .201; nor morality, b = 0.13, SE = 0.09, $t(530)$ = 1.42, p = .155, and there were no significant interactions with familiarity for either measure.

Table 4. Means, SDs, and inferential tests for Study D

Outcome	**Intelligence Condition**					**Morality Condition**				
	Baseline		**Post-Increase**			**Baseline**		**Post-Increase**		

	Mean	SD	Mean	SD	Test	Mean	SD	Mean	SD	Test
Intelligence	**2.93**	**1.32**	**5.18**	**1.21**	***t*(184) = 19.94, *p* < .001, BF_{10} = 2.7 x 10^{44}, *d* =1.47**	**2.84**	**1.40**	**4.42**	**1.42**	***t*(146) = 12.42, *p* < .001, BF_{10} = 3.1 x 10^{21}, *d* = 1.02**
Intelligence (Comparison)	**-1.48**	**1.45**	**0.45**	**1.54**	***t*(184) = 18.33, *p* < .001, BF_{10} = 1.0 x 10^{40}, *d* = 1.35**	**-1.61**	**1.42**	**0.00**	**1.35**	***t*(146) = 13.94, *p* < .001, BF_{10} = 2.7 x 10^{25}, *d* = 1.15**
Morality	**2.68**	**1.41**	**3.12**	**1.27**	***t*(184) = 5.13, *p* < .001, BF_{10} = 1.5 x 10^{4}, *d* = 0.38**	**2.49**	**1.46**	**4.67**	**1.46**	***t*(146) = 16.37, *p* < .001, BF_{10} = 3.7 x 10^{31}, *d* = 1.35**
Morality (Comparison)	**-1.55**	**1.26**	**-1.11**	**1.27**	***t*(184) = 6.10, *p* < .001, BF_{10} = 1.5 x 10^{6}, *d* = 0.45**	**-1.65**	**1.39**	**-0.11**	**1.43**	***t*(146) = 14.23, *p* < .001, BF_{10} = 1.5 x 10^{26}, *d* = 1.17**
Trust	**2.96**	**1.33**	**3.37**	**1.39**	***t*(184) = 4.97, *p* < .001, BF_{10} = 7.4 × 10^{3}, *d* = 0.37**	**3.19**	**1.40**	**3.98**	**1.34**	***t*(146) = 6.95, *p* < .001, BF_{10} = 7.6 × 10^{7}, *d* = 0.57**
Danger	**3.92**	**1.54**	**4.81**	**1.57**	***t*(184) = 9.19, *p* < .001, BF_{10} = 6.6 × 10^{13}, *d* = 0.68**	**3.90**	**1.75**	**4.05**	**1.70**	***t*(146) = 1.39, *p* = .166, BF_{10} = .24, *d* = 0.11**

5. Study 1a

5.1 Method

5.1.1 Participants

We recruited 850 participants (695 after excluding those failing the pre-registered attention checks) from the United Kingdom (aged between 18 and 83, M_{age} = 41; 411 male, 273 female, 8 non-binary/other) from Prolific Academic in exchange for payment of £0.75 GPB. Using G*power, we determined a sample size of 787 participants for a power of 80% for detecting small effects sizes (d = 0.1). To account for some participants potentially failing attention checks, we recruited 850 in total to maintain the required power. The focal analysis for the power calculation was 3-way interaction predicting intelligence/morality ratings. While we did not meet this number of participants in this study, a post-hoc sensitivity analysis revealed that, due to rounding and the approximations involved in power analysis, we still had 80% power to detect an effect of d = 0.1, due to rounding and the approximations involved in power analysis.

5.1.2 Measures

The measures in this study were slightly modified compared to the pilot studies. First, the morality and intelligence questions were simplified so as not to describe what these features are, ensuring consistency across conditions e.g. "*How intelligent do you think [Tom/Omega AI] is?*". Secondly, the "human comparison" scales are now worded as "average person comparison" scales to make them read more naturally for participants in the human condition. Third, in this study (but not in the future ones) the question about AI familiarity was only asked for participants in the AI conditions.

5.1.3 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at https://osf.io/c7vu2/?view_only=624c12eb066f4b8cbb4d782cc22f1f2c

5.2 Results

Our manipulations were successful, with higher perceived intelligence in the high intelligence conditions for both humans and AI, and higher perceived morality in the high morality conditions for both humans and AI, across both the general question and the question asking in comparison to the average person (See Table 5).

To assess whether the intelligence manipulation affected expected morality, we calculated a pre-registered linear model looking at the effect of intelligence level (high vs. low) and agent (human vs. AI) on ratings of morality, once for the general ratings and once for question on morality in comparison to the average person. For the general ratings, we found that those in the high intelligence condition gave significantly higher morality ratings, $F(1,320) = 39.95$, $p < .001$, $\eta_p^2 = 0.11$, and rated the human agent as being more moral than the AI, $F(1,320) = 155.49$, $p < .001$, $\eta_p^2 = 0.33$, but with no interaction effect, $F(1,320) = 1.96$, $p = .16$, $\eta_p^2 = 0.01$. This was also the case for average-person-comparison ratings, with higher ratings of morality in the high intelligence condition, $F(1,320) = 32.30$, $p < .001$, $\eta_p^2 = 0.09$; and higher ratings

of morality for humans than AI, $F(1,320) = 112.82$, $p < .001$, $\eta_p^2 = 0.26$, but with no interaction effect, $F(1,320) = 0.01$, $p = .916$, $\eta_p^2 = 0.00$. Once again, we again found that while describing an AI as having high intelligence led to higher ratings of morality than for an AI with low intelligence, even the AI with high intelligence was seen as less moral than the average human (M = -0.76, SD = 1.50).

Next, we looked at how the morality manipulation affected expected intelligence in a pre-registered linear model looking at the effect of morality level (high vs. low) and agent (human vs. AI) on ratings of intelligence, once for the general ratings and once for question in comparison to the average person. For the general ratings, alongside main effects whereby those in the high morality condition gave significantly higher intelligence ratings, $F(1,367) = 64.22$, $p < .001$, $\eta_p^2 = 0.15$, and rated the AI agent as being more intelligent than the human, $F(1,367) = 13.48$, $p < .001$, $\eta_p^2 = 0.04$. Additionally, there was an interaction effect, $F(1,367) = 10.40$, $p = .001$, $\eta_p^2 = 0.03$, whereby the difference in rated intelligence between the human and AI was stronger in the low morality condition with the low-morality human seen as particularly unintelligent, $t(175.42) = 4.84$, $p < .001$, $d = 0.71$. This was also the case for average-person-comparison ratings, with higher ratings of intelligence in the high morality condition, $F(1,367) = 49.17$, $p < .001$, $\eta_p^2 = 0.12$; and higher ratings of intelligence compared to average person for the AI than the human, $F(1,367) = 20.02$, $p < .001$, $\eta_p^2 = 0.05$, and again an interaction effect, $F(1,320) = 9.18$, $p = .003$, $\eta_p^2 = 0.02$ whereby, all agents were seen as more intelligent than the average person except the human described as low in morality.

As in previous studies and as predicted, then, we found that participants who read about the high intelligence agents also perceived them as more moral on both measures, and participants who read about the high morality agents also perceived them as more intelligent across both measures. Overall, humans were rated as having higher morality but AI was rated as having higher intelligence and, as in previous studies, while describing an AI as being more intelligent led participants to think the AI was also more moral, even the highly intelligent AI was still rated as less moral than the average person.

Table 5. Means, SDs, and inferential tests in Study 3.

Outcome	**AI Agent Conditions**									
	Intelligence Condition					**Morality Condition**				
	Low		**High**			**Low**		**High**		
	Mean	**SD**	**Mean**	**SD**	**Test**	**Mean**	**SD**	**Mean**	**SD**	**Test**
Intelligence	**2.25**	**1.10**	**6.24**	**0.90**	$t(135.00) = -23.85$, $p < .001$, $BF_{10} = 2.53 \times 10^{48}$, $d = 3.97$	**4.69**	**1.40**	**5.30**	**1.42**	$t(197.55) = -3.06$, $p = .003$, $BF_{10} = 1.14 \times 10^{1}$, $d = 0.43$
Intelligence compared to average person	**-1.70**	**1.18**	**1.55**	**1.38**	$t(142.49) = -15.36$, $p < .001$, $BF_{10} = 3.24 \times 10^{28}$, $d = 2.53$	**0.74**	**1.62**	**1.31**	**1.60**	$t(196.98) = -2.50$, $p = .013$, $BF_{10} = 2.80 \times 10^{0}$, $d = 0.35$
Morality	**2.55**	**1.18**	**3.49**	**1.29**	$t(143.86) = -4.62$, $p < .001$, $BF_{10} = 1.91 \times 10^{3}$, $d = 0.76$	**2.76**	**1.48**	**5.86**	**1.24**	$t(187.63) = -16.04$, $p < .001$, $BF_{10} = 2.54 \times 10^{34}$, $d = 2.28$
Morality compared to average person	**-1.51**	**1.23**	**-0.76**	**1.50**	$t(141.23) = -3.30$, $p = .001$, $BF_{10} = 2.25 \times 10^{1}$, $d = 0.54$	**-1.59**	**1.38**	**0.79**	**1.46**	$t(198.00) = -11.84$, $p < .001$, $BF_{10} = 2.97 \times 10^{21}$, $d = 1.67$
Trust	**2.70**	**1.34**	**4.29**	**1.41**	$t(144.00) = -6.99$, $p < .001$, $BF_{10} = 7.81 \times 10^{7}$, $d = 1.16$	**3.1**	**1.42**	**4.68**	**1.3**	$t(193.84) = -8.18$, $p < .001$, $BF_{10} = 1.93 \times 10^{11}$, $d = 1.16$
Danger	**4.38**	**1.58**	**4.39**	**1.53**	$t(142.95) = -0.02$, $p = .980$, $BF_{10} = 1.78 \times 10^{-1}$, $d = 0.00$	**4.69**	**1.62**	**4.01**	**1.35**	$t(187.25) = 3.23$, $p = .001$, $BF_{10} = 1.95 \times 10^{1}$, $d = 0.46$
Outcome	**Human Agent Conditions**									
	Intelligence Condition					**Morality Condition**				
	Low		**High**			**Low**		**High**		

	Mean	SD	Mean	SD	Test	Mean	SD	Mean	SD	Test
Intelligence	**2.36**	**1.03**	**6.65**	**0.70**	$t(176.00) = -33.30$, $p < .001$, $BF_{10} = 2.04 \times 10^{69}$, $d = 4.74$	**3.81**	**1.04**	**5.24**	**0.87**	$t(162.97) = -9.77$, $p < .001$, $BF_{10} = 1.19 \times 10^{15}$, $d = 1.50$
Intelligence compared to average person	**-1.61**	**1.07**	**2.39**	**0.72**	$t(176.00) = -29.86$, $p < .001$, $BF_{10} = 2.84 \times 10^{62}$, $d = 4.25$	**-0.33**	**1.14**	**1.10**	**0.91**	$t(160.25) = -9.10$, $p < .001$, $BF_{10} = 2.08 \times 10^{13}$, $d = 1.39$
Morality	**4.25**	**0.88**	**4.85**	**1.03**	$t(136.42) = -4.05$, $p < .001$, $BF_{10} = 4.11 \times 10^{2}$, $d = 0.64$	**1.99**	**1.20**	**6.50**	**0.81**	$t(146.98) = -28.80$, $p < .001$, $BF_{10} = 2.25 \times 10^{63}$, $d = 4.42$
Morality compared to average person	**-0.12**	**0.81**	**0.60**	**1.06**	$t(125.80) = -4.88$, $p < .001$, $BF_{10} = 1.78 \times 10^{4}$, $d = 0.78$	**-2.04**	**1.04**	**2.17**	**0.91**	$t(165.54) = -28.15$, $p < .001$, $BF_{10} = 7.79 \times 10^{61}$, $d = 4.31$
Trust	**4.14**	**1.03**	**4.81**	**1.12**	$t(143.52) = -4.01$, $p < .001$, $BF_{10} = 2.91 \times 10^{2}$, $d = 0.62$	**2.15**	**1.25**	**6.15**	**0.87**	$t(150.29) = -24.22$, $p < .001$, $BF_{10} = 1.55 \times 10^{53}$, $d = 3.71$
Danger	**3.18**	**1.26**	**3.40**	**1.40**	$t(141.85) = -1.09$, $p = .279$, $BF_{10} = 2.92 \times 10^{-1}$, $d = 0.17$	**4.54**	**1.48**	**1.98**	**0.92**	$t(140.41) = 13.61$, $p < .001$, $BF_{10} = 4.83 \times 10^{25}$, $d = 2.09$

Using pre-registered linear models, we next assessed the extent to which the intelligence and morality manipulations affected perceptions of trust and danger across humans and AI targets. For trust, we found a significant three-way interaction between trait (intelligence vs. morality), level (high vs. low), and agent (human vs. AI) whereby both higher intelligence and morality was associated with more trust, but the effect of morality was particularly strong for the human agent with the low-morality human being seen as less trustworthy than the low-morality AI, and the high morality human seen as more trustworthy than the high morality AI. For perceived danger, we again found a significant three-way interaction in which higher morality was associated with reduced danger for both agents, but the intelligence manipulation only influenced reduced fear of the human, with no effect of intelligence manipulation on perceived danger from the AI (see Table 6).

As in previous studies, across conditions we found significant correlations between morality and intelligence on both general expectations, $r(693) = 0.34$, $p < .001$, and average-person-comparison $r(693) = 0.33$, $p < .001$. Looking at correlations within each condition, there was a significant correlation between ratings of intelligence and morality for six of the eight conditions, with the exceptions of the high-morality AI condition, $r(101) = 0.18$, $p = .069$, and the low-morality AI condition, $r(95) = 0.16$, $p = .117$.

To explore the effects of AI familiarity on perceived morality and intelligence we added self-reported familiarity (centered) into mixed-models looking at the effects of level and trait type (but only for the AI conditions, since in this study we only asked for participants in the AI conditions). Self-reported AI familiarity had no significant main effect on ratings of intelligence, $b = 0.04$, $SE = 0.09$, $t(338) = 0.38$, $p = .702$, and morality, $b = -0.15$, $SE = 0.10$, $t(338) = -1.55$, $p = .123$, and there were no significant interactions of level or trait type with familiarity for either measure.

Table 6. Model results for the effects of condition on trust and danger in Study 3.

		Trust			Danger		
	df	***F***	***p***	η_p^2	***F***	***p***	η_p^2
Level	**1**	**432.44**	**<.001**	**0.39**	**49.29**	**<.001**	**0.07**
Agent	**1**	**43.09**	**<.001**	**0.06**	**103.39**	**<.001**	**0.13**
Trait	**1**	**0.14**	**0.706**	**0.00**	**0.09**	**0.761**	**0.00**
Level × Agent	**1**	**15.81**	**<.001**	**0.02**	**15.05**	**<.001**	**0.02**
Level × Trait	**1**	**77.86**	**<.001**	**0.1**	**65.47**	**<.001**	**0.09**
Agent × Trait	**1**	**14.39**	**<.001**	**0.02**	**0.00**	**0.996**	**0.00**
Level × Agent × Trait	**1**	**79.05**	**<.001**	**0.1**	**23.92**	**<.001**	**0.03**

Residuals	**687**						
***p < 0.05; **p < 0.01; ***p < 0.001**							

6. Study 1b

6.1 Method

6.1.1 Participants

We recruited 850 participants (739 after excluding those failing the pre-registered attention checks) from the United Kingdom (aged between 18 and 86, M_{age} = 44; 338 male and 396 female, 2 non-binary/other) from Prolific Academic in exchange for payment of £0.75 GPB. Power analysis was conducted via G*Power 3.1.9.7. The focal analysis was 3-way interaction Rating ~ Trait * Condition * Agent. We determined a Power estimate of 80% for small effects sizes (f = 0.1) estimated around 787 participants, and oversampled to recruit 850 to account for attention checks. While our final sample size was lower than indicated by our power analysis, due to rounding and the approximations involved in power analysis, we still had 80% power to detect an effect of f = 0.1.

6.1.2 Measures

Measures are the same as previous studies with the addition of a more objective measure regarding the participants' knowledge of AI, which we took from PEW. Example questions include: Thinking about customer service, which of the following uses artificial intelligence (AI)? (Answers: *"A detailed Frequently Asked Questions webpage"; "An online survey sent to customers that allows them to provide feedback"; "A contact page with a form available to customers to provide feedback; A chatbot that immediately answers customer questions"; "Not sure"*).

6.1.3 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at https://osf.io/w34g7/?view_only=6f4ec1fde7ee413fa626c2671b0353b7

6.2 Results

Our manipulations were successful. To check the effects of our augmentations, we calculated a pre-registered linear model looking at the effect of the intelligence augmentation (pre vs. post) and agent (human vs. AI) on ratings of intelligence (for those in the intelligence conditions) or morality (for those in the morality conditions). We did this once for the general ratings and once for question in comparison to the average person, where for both questions participants first rated perceived intelligence or morality and then again after the intelligence or morality augmentation, "as a result of this new breakthrough". Looking first at the conditions where intelligence was manipulated and ratings of intelligence on the general ratings, we found a significant effect of augmentation, b = 4.19, SE = 0.11, t(742) = 37.66, p < .001 where intelligence ratings were higher post-augmentation, and a significant interaction, b = -0.76, SE = 0.16, t(742) = -4.83, p < .001, whereby the augmentation was thought to change human intelligence more than AI. For the question in comparison to the average person, we also found a main effect of augmentation, b = 3.23, SE = 0.12, t(371) = 27.30, p = .001, but no interaction with agent, b = -0.15, SE = 0.17, t(371) = -0.90, p = .370. Turning then to look at whether participants in the increased morality conditions perceived higher morality post-augmentation, we found a significant main effect of augmentation on morality, b = 3.93, SE = 0.11, t(368) = 34.27, p <.001, and a significant interaction, b = -0.32, SE = 0.16, t(368) = -2.01, p = .045, whereby the augmentation had a slightly stronger effect for the human. For the question on comparison to the average person, the same pattern was found with higher morality scores after the morality augmentation, b = 3.41, SE = 0.11, t(368) = 30.12, p <.001, and an interaction with agent, b = -0.39, SE = 0.16, t(368) = -2.47, p = .014.

The same qualitative patterns from the main text are seen for the ratings of morality in comparison to the average person, with a main effect of augmentation, b = 0.40, SE = 0.10, t(371) = 3.81, p < .001, a main effect of agent, b = -1.41, SE = 0.13, t(664.80) = -10.97, p < .001, and a significant interaction, b = 0.68, SE = 0.15, t(371) = 4.62, p < .001. Notably, while describing an AI as becoming more intelligent led participants to rate the AI as also becoming moral, the highly intelligent AI was still seen as less moral (but more intelligent) than the average person.

This was true also for the intelligence ratings in comparison to the average person, with main effects of augmentation, b = 0.70, SE = 0.10, t(368) = 7.19, p < .001, agent type, b = 0.40, SE = 0.13, t(604.90) = 3.07, p = .002, and a significant interaction between them, b = 0.36, SE = 0.14, t(368) = 2.67, p = .008.

We then looked at whether augmented morality or augmented intelligence had a stronger effect in a pre-registered linear model testing the difference between pre and post morality ratings for the increased intelligence group and pre and post intelligence for the increased morality group. Looking at the general ratings, there was a main effect whereby 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 significant interaction between trait and agent type, b = 0.46, SE = 0.20, t(735) = 2.27, p = .023, whereby the intelligence augmentation had a stronger effect on perceived morality for AI than humans. Looking at the question on comparison to the average person, we found a main effects of augmentation type, b = -0.30, SE = 0.14, t(735) = -2.09, p = .037, whereby increasing morality changed perceptions of intelligence more than increasing intelligence changed morality, and a main effect of agent type, b = 0.36, SE = 0.14, t(735) = -2.09, p = .011, whereby the augmentations changed perceptions of AI more than humans. There was, however, no significant interaction effect, b = 0.32, SE = 0.20, t(735) = 1.59, p = .112.

Table 7. Means, SDs, and inferential tests in Study 4.

Outcome	**AI Agent**			

	Intelligence Augmentation					Morality Augmentation				
	Pre		Post			Pre		Post		
	Mean	SD	Mean	SD	Test	Mean	SD	Mean	SD	Test
Intelligence	2.35	1.27	5.77	1.23	$t(184) = -27.64$, $p < .001$, $BF_{10} = 6.53 \times 10^{63}$, $d = 2.03$	4.26	1.49	5.25	1.25	$t(186) = -9.16$, $p < .001$, $BF_{10} = 5.64 \times 10^{13}$, $d = 0.67$
Intelligence compared to average person	-1.60	1.45	1.48	1.43	$t(184) = -27.64$, $p < .001$, $BF_{10} = 6.53 \times 10^{63}$, $d = 2.03$	-0.02	1.57	1.04	1.37	$t(186) = -10.64$, $p < .001$, $BF_{10} = 7.63 \times 10^{17}$, $d = 0.78$
Morality	2.43	1.34	3.54	1.51	$t(184) = -10.56$, $p < .001$, $BF_{10} = 4.03 \times 10^{17}$, $d = 0.78$	1.89	0.81	5.50	1.49	$t(186) = -30.40$, $p < .001$, $BF_{10} = 2.62 \times 10^{70}$, $d = 2.22$
Morality compared to average person	-1.62	1.35	-0.54	1.61	$t(184) = -9.67$, $p < .001$, $BF_{10} = 1.30 \times 10^{15}$, $d = 0.71$	-2.13	0.88	0.89	1.59	$t(186) = -25.99$, $p < .001$, $BF_{10} = 1.82 \times 10^{60}$, $d = 1.90$
Trust	2.49	1.42	4.29	1.53	$t(184) = -13.94$, $p < .001$, $BF_{10} = 2.43 \times 10^{27}$, $d = 1.02$	2.48	1.30	4.58	1.53	$t(186) = -16.23$, $p < .001$, $BF_{10} = 1.49 \times 10^{34}$, $d = 1.19$
Danger	4.49	1.69	4.51	1.54	$t(184) = -0.11$, $p = .916$, $BF_{10} = 8.25 \times 10^{-2}$, $d = 0.01$	5.39	1.31	4.11	1.52	$t(186) = 9.46$, $p < .001$, $BF_{10} = 3.82 \times 10^{14}$, $d = -0.69$
Outcome	Human Agent									
	Intelligence Augmentation					Morality Augmentation				
	Pre		Post			Pre		Post		
	Mean	SD	Mean	SD	Test	Mean	SD	Mean	SD	Test
Intelligence	2.24	0.83	6.42	0.89	$t(185) = -41.38$, $p < .001$, $BF_{10} = 4.57 \times 10^{91}$, $d = 3.03$	3.61	0.99	4.16	1.09	$t(180) = -6.02$, $p < .001$, $BF_{10} = 9.53 \times 10^{5}$, $d = 0.45$
Intelligence compared to average person	-1.16	0.93	2.06	0.97	$t(185) = -28.77$, $p < .001$, $BF_{10} = 3.93 \times 10^{66}$, $d = 2.11$	-0.42	0.98	0.28	1.01	$t(180) = -7.51$, $p < .001$, $BF_{10} = 2.61 \times 10^{9}$, $d = 0.56$
Morality	3.99	1.05	4.22	0.98	$t(185) = -2.33$, $p = .021$, $BF_{10} = 1.14 \times 10^{0}$, $d = 0.17$	2.21	0.88	6.14	1.20	$t(180) = -36.08$, $p < .001$, $BF_{10} = 3.98 \times 10^{80}$, $d = 2.68$
Morality compared to average person	-0.20	0.92	0.19	0.95	$t(185) = -4.08$, $p < .001$, $BF_{10} = 2.14 \times 10^{2}$, $d = 0.30$	-1.56	0.94	1.86	1.17	$t(180) = -31.37$, $p < .001$, $BF_{10} = 1.68 \times 10^{71}$, $d = 2.33$
Trust	3.99	1.26	4.23	1.14	$t(185) = -2.04$, $p = .043$, $BF_{10} = 6.25 \times 10^{-}$	2.11	1.02	4.99	1.45	$t(180) = -22.27$, $p < .001$, $BF_{10} = 1.29 \times$

					^{1}, d = 0.15					10^{50}, d = 1.66
Danger	**3.34**	**1.36**	**3.98**	**1.52**	$t(185) = -6.36$, $p < .001$, $BF_{10} = 5.47 \times 10^{6}$, $d = 0.47$	**4.96**	**1.35**	**3.15**	**1.47**	$t(180) = 13.14$, $p < .001$, $BF_{10} = 7.69 \times 10^{24}$, $d = -0.98$

Table 8. Effects of augmentation (pre vs. post), augmented trait (morality vs intelligence), and agent (human vs AI) on perceptions of trustworthiness and danger.

	Trust	Danger
Augmentation [Post]	$b = 2.88$, $SE = 0.13$, $t(739) = 22.67$, $p < .001$	$b = -1.81$, $SE = 0.13$, $t(739) = -13.47$, $p < .001$
Trait [Intelligence]	$b = 1.88$, $SE = 0.14$, $t(1428.6) = 13.41$, $p < .001$	$b = -1.62$, $SE = 0.15$, $t(1392.6) = -10.53$, $p < .001$
Agent [AI]	$b = 0.37$, $SE = 0.14$, $t(1428.6) = 2.61$, $p = .009$	$b = 0.43$, $SE = 0.15$, $t(1392.6) = 2.83$, $p = .005$
Augmentation × Trait	$b = -2.65$, $SE = 0.18$, $t(739) = -14.81$, $p < .001$	$b = 2.45$, $SE = 0.19$, $t(739) = 13.01$, $p < .001$
Augmentation × Agent	$b = -0.78$, $SE = 0.18$, $t(739) = -4.35$, $p < .001$	$b = 0.52$, $SE = 0.19$, $t(739) = 2.78$, $p = .006$
Augmentation × Agent	$b = -1.87$, $SE = 0.20$, $t(1428.6) = -9.46$, $p < .001$	$b = 0.72$, $SE = 0.22$, $t(1392.6) = 3.32$, $p < .001$
Augmentation × Trait × Agent	$b = 2.35$, $SE = 0.25$, $t(739) = 9.31$, $p < .001$	$b = -1.15$, $SE = 0.27$, $t(739) = -4.34$, $p < .001$

Across conditions and as in previous studies, we found significant correlations between morality and intelligence. Prior to the augmentation across conditions, there was a significant negative correlation between general ratings of intelligence and morality for the human $r(365) = -0.24$, $p < .001$, and no correlation for the AI agent, $r(370) = 0.09$, $p = .096$. After the augmentation, there was a significant positive correlation between intelligence and morality for all conditions (all rs between 26 and .56, all ps < .001: see full results at the OSF). In other words, at baseline people thought an AI who was higher in intelligence would not be higher in morality and that a human who was higher in intelligence would be lower in morality, but after the augmentations people perceived a positive correlation between them.

Lastly, we explored the familiarity-with-AI self-report measure and the objective AI knowledge index using mixed models as in previous studies. There was again no main effect of self-reported AI familiarity (centered) on general ratings of intelligence, $b = 0.10$, $SE = 0.07$, $t(1422.8) = 1.42$, $p = .157$, or morality, $b = -0.03$, $SE = 0.07$, $t(1404.7) = -0.36$, $p = .718$, and no interactions. Looking at the PEW questions as a more objective measure of AI knowledge, the same pattern was observed with no main effect of knowledge on perceptions of intelligence, $b = 0.05$, $SE = 0.06$, $t(1404.7) = 0.84$, $p = .40$, or morality, $b = 0.06$, $SE = 0.07$, $t(1347) = -0.94$, $p = .348$, and no interactions with agent type.

7. Study 2a

7.1 Method

7.1.1 Participants

We recruited 400 participants (353 after excluding those failing the pre-registered from the United Kingdom (aged between 18 and 77, M_{age} = 41, 158 male; 189 female; 3 non-binary/other) from Prolific Academic in exchange for payment of £0.75 GPB. The sample size and corresponding justification matched that of Studies 1 and 2.

7.1.2 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at https://osf.io/7ebw3/?view_only=aab49b38fe1841ebb38efa1c0aa60e05

7.2 Results

Our manipulations were successful, with higher perceived intelligence in the high intelligence conditions, $F(1,349) = 1548.19$, $p < .001$, $\eta_p^2 = 0.82$, and no interaction with agent, $F(1,349) = 3.04$, $p = .082$, $\eta_p^2 = 0.01$. We averaged across moral competence items to create a moral competence index (for overall ratings $\alpha = 0.88$; for average-person-comparison measures $\alpha = 0.90$) and averaged across moral motivations items to create a moral motivation index (for overall ratings, $\alpha = 0.78$; and for average-person-comparison measures $\alpha = 0.77$).

Similar results as the main text were found for the question on moral competence in comparison to the average person, with a main effect of level, $F(1,349) = 327.52$, $p < .001$, $\eta_p^2 = 0.48$, agent, $F(1,349) = 70.24$, $p < .001$, $\eta_p^2 = 0.17$, and a significant interaction $F(1,349) = 8.64$, $p = .004$, $\eta_p^2 = 0.02$ with the same pattern. We observed that both human and AI agents described as low in intelligence were seen as less morally competent than the average person, and a human high in intelligence was more morally competent than the average person, but an AI high in intelligence was still seen as only equivalent to the average human person in moral competence.

Similarly for the question on moral motivation in comparison to the average person, there was no interaction effect, $F(1,349) = 2.88$, $p = .091$, $\eta_p^2 = 0.01$, only main effects of level, $F(1,349) = 68.13$, $p < .001$, $\eta_p^2 = 0.16$, and agent, $F(1,349) = 26.71$, $p < .001$, $\eta_p^2 = 0.07$. While the low intelligent AI was seen as having worse moral motivation than the average person, the high intelligence AI was perceived to have as much moral motivation as the average person.

We next explored whether manipulated intelligence level led to a stronger effect on participants ratings of moral competence than moral motivation, and whether this depended on agent, in an exploratory (not pre-registered) mixed model. For the general ratings, we observed a significant three-way interaction between intelligence level, agent, and moral trait, $b = 1.05$, $SE = 0.22$, $t(349) = 4.83$, $p < .001$. Decomposing this, we found a significant interaction between intelligence level and moral trait type on ratings for both AI, $b = -0.67$, $SE = 0.16$, $t(164) = -4.22$, $p < .001$, and humans, $b = -1.72$, $SE = 0.07$, $t(185) = -11.59$, $p < .001$, whereby while high intelligence increased both moral competence and moral motivation for both humans and AI, the effect of the intelligence manipulation on moral competence was predictably stronger than on moral motivation. Moreover, we found that while there was no interaction between agent and moral trait for agents described as low intelligence, $b = -0.05$, $SE = 0.15$, $t(183) = -0.32$, $p = .751$, there was for agents high in intelligence, $b = 1.00$, $SE = 0.16$, $t(166) = 6.24$, $p < .001$ such that there was a stronger difference between moral competence in highly intelligent humans and AI than there was for moral motivation. For the comparison questions, we observed the same pattern of results with a significant three-way interaction, $b = 1.17$, $SE = 0.22$, $t(349) = 5.46$, $p < .001$.

Across conditions there was a significant correlation between ratings of intelligence and both moral competence, $r(351) = 0.70$, $p < .001$, and moral motivation, $r(351) = 0.46$, $p < .001$. Looking by conditions, intelligence and both moral competence and moral motivation were significantly correlated for both humans and AI in the low intelligence conditions, but not the high intelligence conditions.

Finally looking at the effects of AI familiarity and knowledge on the PEW questions in exploratory analyses, we again found little evidence of participants' experience with AI influencing these effects. There were no main effects of either self-reported familiarity or PEW AI knowledge on intelligence, moral competence, nor moral motivation, and there were no interaction effects with age

Table 9. Means, SDs, and inferential tests in Study 5.

Outcome	**AI Agent**					**Human Agent**				
	Intelligence Condition					**Intelligence Condition**				
	Low		**High**			**Low**		**High**		
	Mean	**SD**	**Mean**	**SD**	**Test**	**Mean**	**SD**	**Mean**	**SD**	**Test**
Intelligence	**2.16**	**1.19**	**6.25**	**0.83**	$t(142.35) = -25.58$, $p < .001$, $BF_{10} = 6.31 \times 10^{55}$, $d = -4.01$	**2.12**	**1.09**	**6.59**	**0.91**	$t(184.51) = -30.57$, $p < .001$, $BF_{10} = 1.06 \times 10^{69}$, $d = -4.41$
Intelligence compared to average person	**-1.89**	**1.22**	**2.14**	**1.09**	$t(159.82) = -22.33$, $p < .001$, $BF_{10} = 1.07 \times 10^{48}$, $d = -3.48$	**-2.14**	**0.86**	**2.63**	**0.69**	$t(184.99) = -41.90$, $p < .001$, $BF_{10} = 3.20 \times 10^{90}$, $d = -6.02$
Moral competence	**2.08**	**0.92**	**4.01**	**1.49**	$t(140.66) = -10.08$, $p < .001$, $BF_{10} = 2.96 \times 10^{15}$, $d = -1.55$	**3.05**	**1.13**	**5.64**	**0.91**	$t(184.93) = -17.36$, $p < .001$, $BF_{10} = 9.90 \times 10^{35}$, $d = -2.50$
Moral competence compared to average person	**-1.96**	**1.04**	**-0.04**	**1.62**	$t(144.28) = -9.16$, $p < .001$, $BF_{10} = 1.30 \times 10^{13}$, $d = -1.41$	**-1.27**	**1.04**	**1.40**	**0.95**	$t(181.20) = -18.31$, $p < .001$, $BF_{10} = 2.14 \times 10^{39}$, $d = -2.67$
Moral motivation	**2.77**	**1.41**	**4.03**	**1.40**	$t(163.49) = -5.75$, $p < .001$, $BF_{10} = 2.71 \times 10^{5}$, $d = -0.89$	**3.79**	**0.98**	**4.65**	**1.00**	$t(174.86) = -5.96$, $p < .001$, $BF_{10} = 8.82 \times 10^{5}$, $d = -0.88$
Moral motivation compared to average person	**-1.28**	**1.38**	**-0.03**	**1.48**	$t(163.90) = -5.64$, $p < .001$, $BF_{10} = 1.54 \times 10^{5}$, $d = -0.87$	**-0.42**	**0.91**	**0.41**	**0.87**	$t(179.12) = -6.33$, $p < .001$, $BF_{10} = 4.30 \times 10^{6}$, $d = -0.93$
Trust	**2.51**	**1.43**	**4.48**	**1.55**	$t(163.87) = -8.54$, $p < .001$, $BF_{10} = 5.56 \times 10^{11}$, $d = -1.32$	**4.04**	**1.10**	**4.61**	**1.16**	$t(171.57) = -3.46$, $p < .001$, $BF_{10} = 4.03 \times 10^{1}$, $d = -0.51$
Danger	**4.60**	**1.58**	**4.49**	**1.34**	$t(157.14) =$	**3.65**	**1.17**	**4.10**	**1.48**	$t(154.05) =$

					0.49, p = .627, BF_{10} = 1.87 × 10^{-1}, d = 0.08					**−2.23, p = .028, BF_{10} = 1.78 × 10^{0}, d = −0.34**

8. Study 2b

8.1 Method

8.1.1 Participants

We recruited 400 participants (362 after excluding those failing the pre-registered attention checks) from the United Kingdom (aged between 19 and 82, M_{age} = 42, 145 male, 209 female, 6 non-binary/other) from Prolific Academic in exchange for payment of £0.75 GPB. The sample size matched Studies 2ab, and followed the same sample size justification.

8.1.2 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at https://osf.io/tg86d?view_only=e57fcbf97daf4606afd4028f774dc997

8.2 Results

Our manipulations were successful, with higher perceived intelligence after the intelligence augmentation, b = 4.26, SE = 0.11, t(724) = 37.35, p < .001. We observed a significant interaction between intelligence augmentation and agent, b = -0.55, SE = 0.17, t(724) = -3.30, p < .001, such that the AI was seen as being more intelligent than the human pre-augmentation, but the human more intelligent than the AI post-augmentation.

The same pattern as the main text was observed for the question on moral competence in comparison to the average person, with main effects of augmentation, b = 2.45, SE = 0.11, t(362) = 21.95, p < .001, and agent, b = -0.80, SE = 0.12, t(715.7) = -6.54, p < .001, with no interaction, b = -0.22, SE = 0.16, t(362) = -1.38, p = .169.

Similarly, the same pattern was observed for the question on moral motivation in comparison to the average person, with main effects of augmentation, b = 0.90, SE = 0.11, t(362) = 8.01, p < .001, and agent type, b = -0.92, SE = 0.12, t(720.2) = -7.62, p < .001, and a significant interaction, b = 0.70, SE = 0.16, t(362) = 4.26, p < .001.

Turning to correlation analyses, we observed significant correlations pre-augmentation between intelligence and both moral competence, r(360) = 0.50, p < .001, and moral motivation, r(360) = 0.34, p < .001, with the correlations being significant for each condition. Similarly, there were significant correlations post-augmentation between intelligence and both moral competence, r(360) = 0.47, p < .001, and moral motivation, r(360) = 0.27, p < .001, with the correlations significant for each condition except for that between intelligence and moral motivation for the human agent.

Finally, we conducted the exploratory analysis looking at the effects of AI familiarity and knowledge on the PEW questions. We again found little evidence of participants' experience with AI influencing these effects. There were no main effects of self-reported familiarity on intelligence, moral competence, or moral motivation, and there were no interaction effects with agent. For the PEW AI knowledge, the majority of participants got all the questions correct, leading to singularity fit problems with the analysis, though for completeness we report them. There was a significant main effect of AI knowledge on perceived intelligence whereby people with more AI knowledge rated all agents as less intelligent , b = -0.17, SE = 0.06, t(724) = -2.97, p = .003, though there were no interactions with agent type. The same pattern was seen for perceived moral competence, b = -0.15, SE = 0.06, t(721.3) = -2.49, p = .013, though there was no effect on moral motivation, b = -0.09, SE = 0.06, t(712.4) = -1.57, p = .117

Table 10. Means, SDs, and inferential tests in Study 6.

Outcome	**AI Agent**					**Human Agent**				
	Augmentation Condition					**Augmentation Condition**				
	Pre		**Post**			**Pre**		**Post**		
	Mean	**SD**	**Mean**	**SD**	**Test**	**Mean**	**SD**	**Mean**	**SD**	**Test**
Intelligence	**2.29**	**1.45**	**6.00**	**1.20**	**t(169) = 25.93, p < .001, BF_{10} = 1.72 × 10^{57}, d = 1.99**	**2.19**	**1.00**	**6.45**	**0.78**	**t(191) = 38.50, p < .001, BF_{10} = 1.54 × 10^{88}, d = 2.78**
Intelligence compared to average person	**-1.55**	**1.49**	**1.89**	**1.34**	**t(169) = 23.02, p < .001, BF_{10} = 3.08 × 10^{50}, d = 1.77**	**-1.82**	**0.98**	**2.47**	**0.82**	**t(191) = 42.29, p < .001, BF_{10} = 1.29 × 10^{95}, d = 3.05**
Moral competence	**2.22**	**1.17**	**4.73**	**1.47**	**t(169) = 19.82, p < .001, BF_{10} =**	**3.22**	**0.98**	**5.63**	**1.03**	**t(191) = 21.40, p < .001, BF_{10} =**

					3.51×10^{42}, $d = 1.52$					1.38×10^{49}, $d = 1.54$
Moral competence compared to average person	**-1.81**	**1.17**	**0.41**	**1.49**	$t(169) = 18.24$, $p < .001$, $BF_{10} = 2.57 \times 10^{38}$, $d = 1.40$	**-1.02**	**0.88**	**1.43**	**1.06**	$t(191) = 22.42$, $p < .001$, $BF_{10} = 7.66 \times 10^{51}$, $d = 1.62$
Moral motivation	**2.76**	**1.26**	**4.51**	**1.49**	$t(169) = 13.96$, $p < .001$, $BF_{10} = 5.53 \times 10^{26}$, $d = 1.07$	**3.94**	**0.92**	**4.78**	**1.04**	$t(191) = 7.75$, $p < .001$, $BF_{10} = 1.28 \times 10^{10}$, $d = 0.56$
Moral motivation compared to average person	**-1.26**	**1.25**	**0.35**	**1.48**	$t(169) = 12.54$, $p < .001$, $BF_{10} = 5.69 \times 10^{22}$, $d = 0.96$	**-0.34**	**0.81**	**0.57**	**1.02**	$t(191) = 8.51$, $p < .001$, $BF_{10} = 1.21 \times 10^{12}$, $d = 0.61$
Trust	**2.48**	**1.45**	**4.52**	**1.57**	$t(169) = 13.73$, $p < .001$, $BF_{10} = 1.26 \times 10^{26}$, $d = 1.05$	**4.29**	**1.12**	**4.48**	**1.17**	$t(191) = 1.65$, $p = .100$, $BF_{10} = 3.06 \times 10^{-1}$, $d = 0.12$
Danger	**4.68**	**1.64**	**4.31**	**1.74**	$t(169) = -2.25$, $p = .026$, $BF_{10} = 9.84 \times 10^{-1}$, $d = -0.17$	**3.33**	**1.30**	**3.87**	**1.50**	$t(191) = 4.08$, $p < .001$, $BF_{10} = 2.07 \times 10^{2}$, $d = 0.29$

9. Study 3

9.1 Method

9.1.1 Participants

We recruited 400 participants (365 after excluding those failing the pre-registered attention checks) from the United Kingdom (aged between 18 and 83, M_{age} = 41, 129 male, 228 female, 7 non-binary/other) from Prolific Academic in exchange for payment of £0.75 GPB. Sample size matches the previous studies.

9.1.2 Transparency and Openness

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/vsg6u/?view_only=b49015d1d2c9459a99aab41777fff76e. This study was pre-registered at https://osf.io/raxv2/?view_only=a151313ce7bf4718b8e7eab9842a8fa6.

9.2 Results

Our description of the intelligence augmentation in vignette was successful, with a one-sample t-test looking at how people thought the intelligence augmentation would affect the agent's intelligence showing that overall, people thought the agents would have their intelligence increased as a result of the augmentation, as described, $t(364) = 71.55$, $p < .001$, $BF_{10} = 1.90 \times 10^{212}$, $d = 3.74$. This was observed for both the AI, $t(186) = 60.90$, $p < .001$, $BF_{10} = 6.71 \times 10^{120}$, $d = 4.45$, and the human, $t(177) = 43.38$, $p < .001$, $BF_{10} = 2.69 \times 10^{92}$, $d = 3.25$.

We then turned to look at how the expected degree of change in moral competence and moral motivation predicted trust. In an exploratory analysis entering moral competence and moral motivation (centered) as predictors in separate models on trust, we found that moral motivation predicted increased trust, $b = 0.51$, $SE = 0.10$, $t(361) = 5.07$, $p < .001$, with no interaction with agent type, $b = 0.18$, $SE = 0.13$, $t(361) = 1.36$, $p = .175$, while moral competence had no main effect on trust, $b = 0.08$, $SE = 0.09$, $t(361) = 0.90$, $p = .370$, but an interaction with agent type, $b = 0.35$, $SE = 0.13$, $t(361) = 2.69$, $p = .007$, such that perceived moral competence predicted increased trust more for AI agents than humans. Taken together, these results suggest that when look at moral competence alone we see a positive effect of trust, seemingly because people expect moral motivation to rise too; but when we hold perceived moral motivation fixed, perceived moral competence reduces trust, presumably because participants see that the AI is getting more morally knowledgeable but not getting more caring.

For danger, in the separate models we found a non-significant but trending negative effect of moral motivation, b = -0.26, $SE = 0.14$, $t(361) = -1.87$, $p = .062$, and no effect of moral competence, b = -0.04, $SE = 0.12$, $t(361) = -0.31$, $p = .756$, with no interactions with agent type for either moral motivation, $b = 0.02$, $SE = 0.18$, $t(361) = 0.13$, $p = .895$, or moral competence, $b = 0.03$, $SE = 0.17$, $t(361) = 0.17$, $p = .86$. When entering the predictors into the same model, as outlined in our pre-registration, we found no main effect of moral competence, $b = 0.12$, $SE = 0.14$, $t(357) = 0.87$, $p = .388$, but only a negative effect of moral motivation predicting reduced danger, b = -0.36, $SE = 0.17$, $t(357) = -2.09$, $p = .038$, with no interactions. Together, these results suggest perceived moral motivation is a more important driver of reduced fear than moral competence.

Looking at correlations, across conditions we found a positive correlation between ratings of intelligence and moral competence, $r(363) = 0.32$, $p < .001$, and intelligence and moral motivations, $r(360) = 0.17$, $p = .002$. The correlations between intelligence and moral competence were significant for both AI, $r(185) = 0.34$, $p < .001$, and human, $r(176) =$

0.31, $p < .001$. For intelligence and moral motivation, the correlations was significant for participants thinking about the AI agent, $r(185) = 0.19$, $p = .010$, but not for the human, $r(176) = 0.14$, $p = .057$.

Finally, we again found little evidence of participants' experience with AI influencing these effects. There were no main effects of self-reported familiarity or AI knowledge on the PEW questions on intelligence, moral competence, or moral motivation, and there were no interaction effects with agent.

Table 11. Means, SDs, and inferential tests in Study 7.

Outcome	Agent Condition						Test
	Human			AI			
	Mean	SD	One Sample Test From Zero	Mean	SD	One Sample Test From Zero	Independent Samples T-Test between AI and Human
Intelligence	41.49	12.76	$t(177) = 43.38$, $p < .001$, $BF_{10} = 2.69 \times 10^{92}$, $d = 3.25$	43.09	9.67	$t(186) = 60.90$, $p < .001$, $BF_{10} = 6.71 \times 10^{120}$, $d = 4.45$	$t(329.74) = -1.34$, $p = .180$, $BF_{10} = 2.78 \times 10^{-1}$, $d = -0.14$
Moral competence	19.19	15.58	$t(177) = 16.43$, $p < .001$, $BF_{10} = 1.33 \times 10^{34}$, $d = 1.23$	19.45	15.58	$t(186) = 17.07$, $p < .001$, $BF_{10} = 3.84 \times 10^{36}$, $d = 1.25$	$t(362.09) = -0.16$, $p = .876$, $BF_{10} = 1.17 \times 10^{-1}$, $d = -0.02$
Moral motivation	11.81	13.25	$t(177) = 11.89$, $p < .001$, $BF_{10} = 1.58 \times 10^{21}$, $d = 0.89$	14.19	15.45	$t(186) = 12.56$, $p < .001$, $BF_{10} = 2.71 \times 10^{23}$, $d = 0.92$	$t(359.19) = -1.58$, $p = .115$, $BF_{10} = 3.81 \times 10^{-1}$, $d = -0.17$
Trust	3.55	15.89	$t(177) = 2.98$, $p = .003$, $BF_{10} = 5.99 \times 10^{0}$, $d = 0.22$	18.05	23.03	$t(186) = 10.72$, $p < .001$, $BF_{10} = 1.28 \times 10^{18}$, $d = 0.78$	$t(331.43) = -7.03$, $p < .001$, $BF_{10} = 4.91 \times 10^{8}$, $d = -0.73$
Danger	4.23	20.84	$t(177) = 2.71$, $p = .007$, $BF_{10} = 2.88 \times 10^{0}$, $d = 0.20$	12.80	27.69	$t(186) = 6.32$, $p < .001$, $BF_{10} = 4.56 \times 10^{6}$, $d = 0.46$	$t(344.85) = -3.35$, $p < .001$, $BF_{10} = 2.24 \times 10^{1}$, $d = -0.35$

10. Deviations from Pre-Registration

With the exception of Study 1a, all studies were pre-registered on the Open Science Framework. We have described in each results section where analyses were pre-registered or non-pregistrered additional analyses, but there were also some minor deviations outlined below.

Study	Deviation	Reason
Pilot Study A & B	We said that we would perform analysis in Jamovi, but have performed them in R using other packages instead.	We have used other R packages for ease of reproducibility. As can be seen in the results output on the OSF, the results from Jamovi are – as one would expect - identical to those through base R.
Pilot Study A & B	The analysis on Trust was listed as exploratory, but we have presented them all the together in the paper in the same table.	We have presented them all the together in the paper because trust is a pre-registered DV in later studies.
Study 1B & 2B	We said that we would conduct ANOVAs, but because we have a mixed design we have instead conducted and reported a linear mixed model.	These are structurally the same analyses but we have reported in regression-style instead of ANOVA-style for clarity.