WORKING PAPERS

N° 1387

November 2022

"Humans Feel Too Special for Machines to Score Their Morals"

Zoe Purcell and Jean-François Bonnefon

Toulouse School of Economics

Humans Feel Too Special for Machines to Score Their Morals

Purcell, Z. A.[1] & Bonnefon, J. F.[1,2]

1. Artificial and Natural Intelligence Toulouse Institute, University of Toulouse, Toulouse, France
2. Toulouse School of Economics and Centre National de la Recherche Scientifique (TSM-R), Toulouse, France

We acknowledge support from grant ANR-19-PI3A-0004, grant ANR-17-EURE-0010, and the research foundation TSE-Partnership.

Abstract

Artificial Intelligence (AI) can be harnessed to create sophisticated social and moral scoring systems — enabling people and organizations to form judgements of others at scale. However, it also poses significant ethical challenges and is, subsequently, the subject of wide debate. As these technologies are developed and governing bodies face regulatory decisions, it is crucial that we understand the attraction or resistance that people have for AI moral scoring. Across four experiments, we show that the acceptability of moral scoring by AI is related to expectations about the quality of those scores, but that expectations about quality are compromised by people's tendency to see themselves as morally peculiar. We demonstrate that people overestimate the peculiarity of their moral profile, believe that AI will neglect this peculiarity, and resist for this reason the introduction of moral scoring by AI.

People feel too special for machines to score their morals

Moral traits such as loyalty and generosity hold a special role in our personal and social identities[1–5] in particular because they help us to build the good reputation that is necessary to thrive as a member of a cooperative society[6–11]. In small-scale societies, traditional channels like personal experience and gossip can be sufficient to acquire information about the moral traits of others[12] – but these channels do not scale up well as societies grow, to the point where we can potentially interact with tens or hundreds of thousands of strangers. One radical solution to this scale problem is to delegate the acquisition and collation of moral information to intelligent machines—in other words, to let Artificial Intelligence (AI) observe the behavior of humans and transform these observations into moral scores intelligible to other humans. Here we show that people are unlikely to accept such AI moral scoring, in large part because they personally feel too special for machines to understand their morals.

We are not concerned in this article with well-accepted practices like the aggregation of peer ratings in online platforms. For example, Uber passengers rate the behavior of their drivers (and the other way round), and these ratings are aggregated in a score that is shown to other customers, in order for these customers to anticipate the behavior of their driver[13–15]. Similar mechanisms exist on other platforms (AirBnB, eBay, etc.), following the same logic: humans rate their experience with another human, and these ratings are aggregated in order for other humans to predict what their experience will be[16–18]. Taken to the extreme, this peer-rating system could generalize to every interaction, as in the famous Black Mirror episode *Nosedive*, which depicts a world in which people rate every encounter they have with each other. But even in this extreme dystopian version, the scoring is done by humans, not by AI.

We are inspired instead by applications where AI observes the online and offline behavioral traces of a human and collates all these traces into one or several moral scores. For example, building on algorithms that extract personality profiles from online activity[19,20], there are now AI tools that attempt to automatically rate the aggressivity or sexism of individual social media users[21–23]. AI moral scoring also fuels large-scale social engineering projects such as the Chinese social credit system, in which a wide range of valued and devalued behaviors are aggregated in a single score for every citizen, with social and legal penalties for citizens who drop beneath a certain score.

Government-operated social credit systems are an extreme form of AI moral scoring and have raised many concerns about abusive social control by authoritarian actors, mischaracterization of individuals, and facilitation of disproportionate social and legal sanctions[24–26]. Indeed, based on the recommendation of its expert group on AI to prohibit mass scale scoring of moral personality (European Commission, 2019[27]), the European Commission is considering a ban on social credit systems operated by its member governments (European Commission, 2021[28]). Even with such a ban in place, though, there would still be room for less extensive forms of AI moral scoring, deployed by public or private actors[29]. For example, people may be willing to let private companies use AI to issue certified scores of their morals, in order to disclose these scores on their CVs or on their dating profiles. As a result, it is important to understand the attraction or resistance that people have for moral scoring by AI, since it will drive their support for AI moral scoring policies and their consumption of AI moral scoring products.

In this article, we seek to show that just as people resist medical AI because they believe AI cannot grasp the peculiarity of their medical profile[30], people will resist moral AI because they think AI cannot grasp the peculiarity of their moral profile. For that purpose, we provide empirical evidence for four claims. First, we show that resistance to AI moral scoring is related to its expected accuracy—in other words, that people are less likely to accept AI moral scoring if they expect AI to mischaracterize their morals

(Claim 1). Second, we show that people overestimate the peculiarity of their moral profile—or more precisely, that they underestimate the prevalence of their moral profile in the population (Claim 2). Third, we show that people believe AI moral scoring to be less accurate for peculiar moral profiles (Claim 3); and fourth, that people believe as a result that AI is likely to mischaracterize their morals (Claim 4). In sum, here we show that people feel too special for AI to score their morals, resulting in resistance to moral scoring by AI.

Results

Study 1

To assess Claim 1 *(people are less likely to accept AI moral scoring if they expect AI to mischaracterize their morals)* we asked a representative UK sample of 446 participants (see Table S1 of the Supplementary Information for demographics) to rate the how acceptable it would be for AI to score each of 17 psychological traits (including 10 moral traits and, for comparison purposes, 4 non-moral personality traits and 3 traits related to mental health), and how good would AI be at scoring each of these traits.

From a descriptive perspective (see Table S2 of the Supplementary Information for detailed statistics), Figure 1A shows that 40% of respondents believed it was (very) unacceptable to use AI to measure moral traits compared to 26% who believed it was (very) acceptable. A linear mixed model with a random intercept for participants showed that this acceptability was higher than that of mental health traits (49% believe it is unacceptable versus 23% acceptable) but lower than that of non-moral personality traits (34% believe it is unacceptable versus 30% acceptable).

Most importantly, Figure 1B provides clear visual evidence of a strong positive relationship between expectations about the quality and acceptability of AI moral scoring, together with an acceptability penalty for negative moral traits. A linear mixed model predicting acceptability ratings from expected quality (with trait and participant as random effect) detected an effect of expected quality on acceptability (B= 0.31, p< .001), and all ratings of acceptability and quality were strongly correlated at the trait level (r between .31 and .65, all p<.001; see Table S2 of the Supplementary Information for detailed statistics).

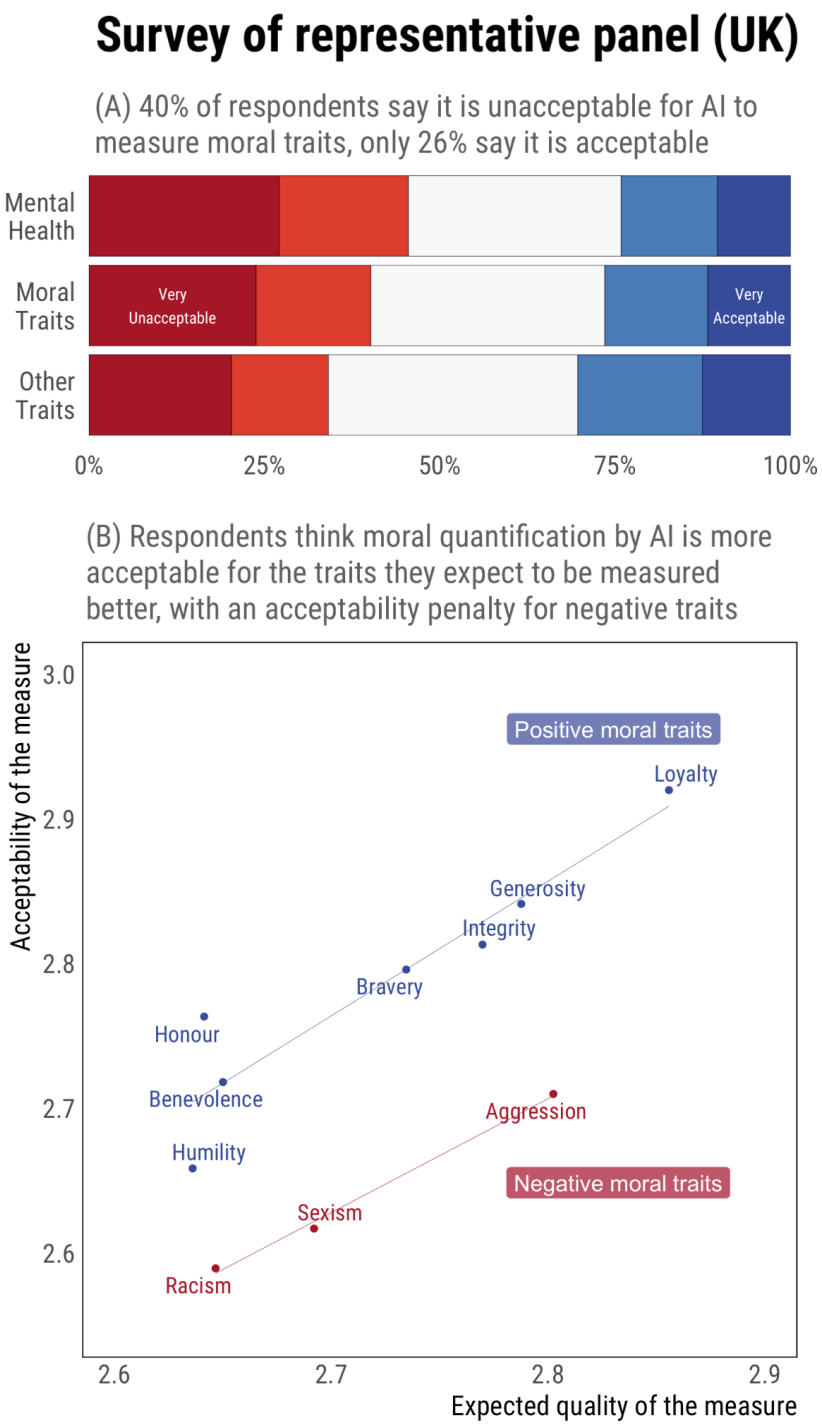


Figure 1. *Results from Study 1 show that (A) participants are more likely to find the scoring of moral traits unacceptable (44%) than acceptable (26%) and that (B) ratings of acceptability are related to perceptions of the quality of the scoring. The acceptability of scoring negative moral traits is lower than the acceptability of scoring positive traits. More details available in Extended Data Figures: Figure E1.*

Studies 2a and 2b

All analyses for Studies 2a and 2b were pre-registered (https://osf.io/x8rgw).

To assess Claim 2 *(people overestimate the peculiarity of their moral profile)*, we had to construct participants' moral profiles, ask them how prevalent they believed their moral profile was, and compare this subjective prevalence to some ground truth about the prevalence of the profile in the population. For this purpose, we used an existing dataset from the *Your Morals* project[31,32] in which 131,015 respondents answered 30 questions about their moral preferences. Using these answers, we categorized the Your Morals respondents into 16 moral profiles corresponding to all combinations of 'low' (below median) and 'high' (above median) values on four moral dimensions (Care, Fairness, Loyalty, Authority), and computed the prevalence of each of these 16 profiles.

Participants in our studies (N = 495 in Study 2a, N = 496 in Study 2b, see Table S1 of the Supplementary Information) answered the same questions as respondents to the Your Morals project, were categorized in one of the 16 same moral profiles, and shown a summary of their moral profile (for an example, see Figure S1 in Section 3 of the Supplementary Information). Participants were then shown an unlabeled histogram of the prevalence of the 16 profiles (see the blue bars in Figure 2) and asked to guess which bar corresponded to their own profile. The only difference between Study 2a and 2b was that their guess was incentivized for accuracy in Study 2b but not in Study 2a. To assess the extent to which participants overestimated the peculiarity of their moral profile, we compared their guess about its prevalence to its actual prevalence.

Figure 1A displays the guesses (red bars) of participants in Study 1A, who were asked which blue bar corresponded to their moral profiles. The distributions of actual and perceived prevalence were markedly different: very few participants believed that their profile was among the most common, and 88% underestimated its prevalence. As pre-registered, we fitted a linear model estimating the difference between actual and perceived prevalence, that included education, gender and politics as predictors. The model's intercept was positive and significantly different from zero (2.08, $p = 0.006$), indicating that people underestimated the prevalence of their moral profile. Demographic covariates had no detectable effect on this underestimation (all *p*s >.40; see Section 3 of the Supplementary Information for detailed results). Study 2b (which offered financial incentives for accurate guesses) delivered very similar findings, as shown in Figure 2B. Once again, the model's intercept was positive and significantly different from zero (B=5.17, $p < .001$) indicating that people underestimated the prevalence of their moral profile. Unlike in Study 2a, the model also detected significant effects of demographic variables. The underestimation effect was larger for males (1.45, $p = 0.040$), and for liberals (1.70, $p = 0.047$; see Section 3 of the Supplementary Information for detailed results).

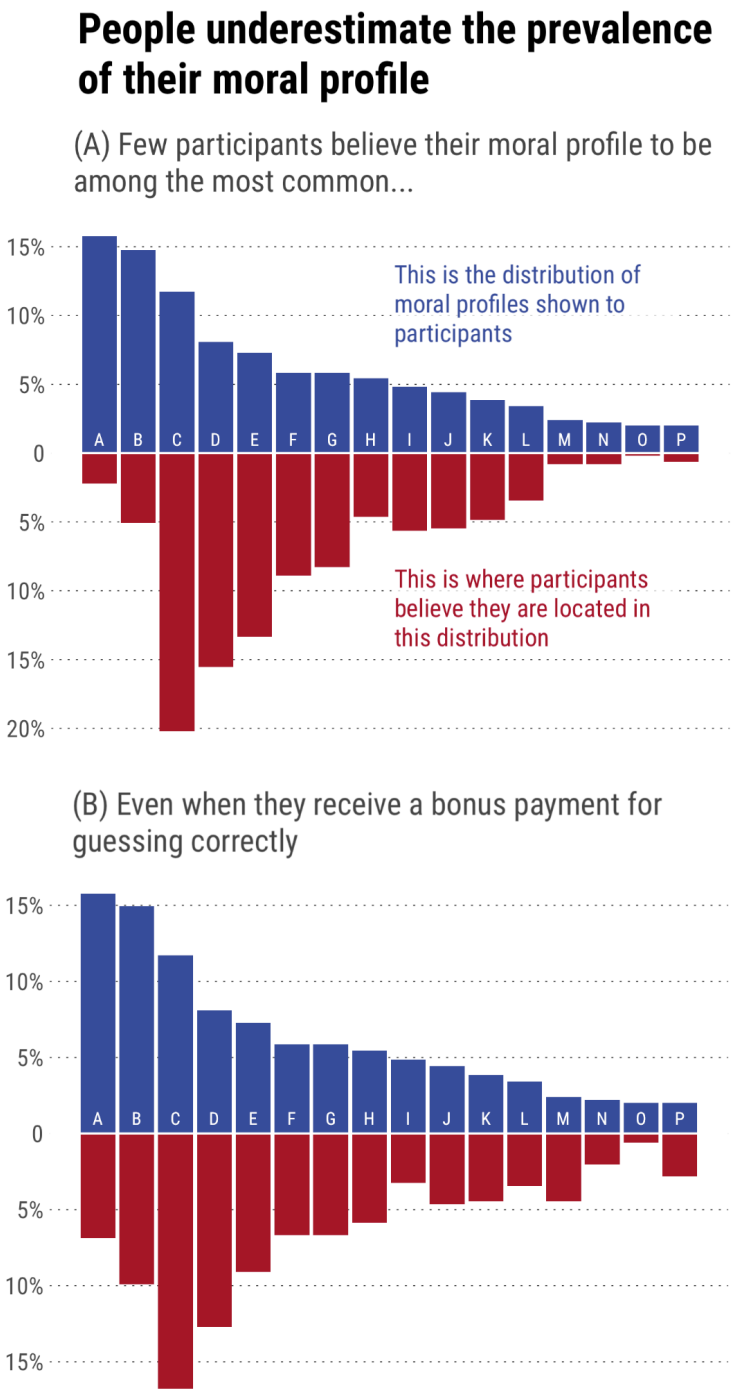


Figure 2. *Results from Studies 2a (A) and 2b (B) show that participants tend to underestimate the prevalence of their moral profile, when shown an unlabeled histogram of the prevalence of the 16 possible moral profiles (blue bars) and asked to guess which bar corresponds to their profile. The red bars display the histogram of guesses by participants. Study 2b offered financial incentives for correct guesses, which did not eliminate the underestimation effect.*

Study 3

All analyses for Study 3 were pre-registered (https://osf.io/x8rgw).

We obtained evidence for Claim 1 *(people are less likely to accept AI moral scoring if they expect AI to mischaracterize their morals)* in Study 1, and for Claim 2 *(people overestimate the peculiarity of their moral profile)* in Studies 2a and 2b. Study 3 pursues three objectives. First, it provides a pre-registered replication of Study 1, whose analyses were not originally pre-registered (see Section 4 of the Supplementary Information for detailed results). Second, it seeks evidence for Claim 3 *(people believe AI moral scoring to be less accurate for peculiar moral profiles)*; and third, it seeks evidence for Claim 4 *(as a result, people believe that AI is likely to mischaracterize their morals)*.

To test Claim 3, we asked 506 participants whether AI moral scoring would be more accurate for typical or unique profiles, whether it would make more errors for typical or unique profiles, whether they would trust AI moral scoring more for typical or unique profiles, and whether they would doubt AI moral scoring more for typical or unique profiles. To test Claim 4, we asked them how well they thought AI would do at scoring their own moral profile.

Participants' responses provided clear support for Claim 3. Figure 3 displays the distribution of average answers to the four performance questions, and skews toward positive values, that is, in the direction of AI moral scoring having trouble with unique profiles. This was confirmed by an intercept-only linear model, which detected that average answers were significantly greater than zero (intercept = 17, SD = 21; $t(5059) = 56.$, $p < .001$).

Participants' responses also provided support for Claim 4, with a twist. As we predicted, the more trouble people believed that AI would have scoring unique profiles, the worse they thought AI would do at scoring their own profile ($B = -0.13$, $p < .001$). However, visual inspection of Figure 3 suggests that this association might only be true for people who at least somewhat believe that AI has more trouble with unique profiles (displayed as red dots in Figure 3), but not true for people who do not hold this belief at all (displayed as blue dots in Figure 3). Since we did not pre-register this prediction, we test it here as an exploratory rather than confirmatory analysis. We fitted two separate linear models testing the blue and red slopes in Figure 3. The blue slope is not significantly different from zero ($B = 0.07$, $p = 0.282$), but the red slope is, and it is steeper than the slope obtained in the pre-registered analysis ($B = -0.18$, $p < .001$; see Section 4 of the Supplementary Information for more details).

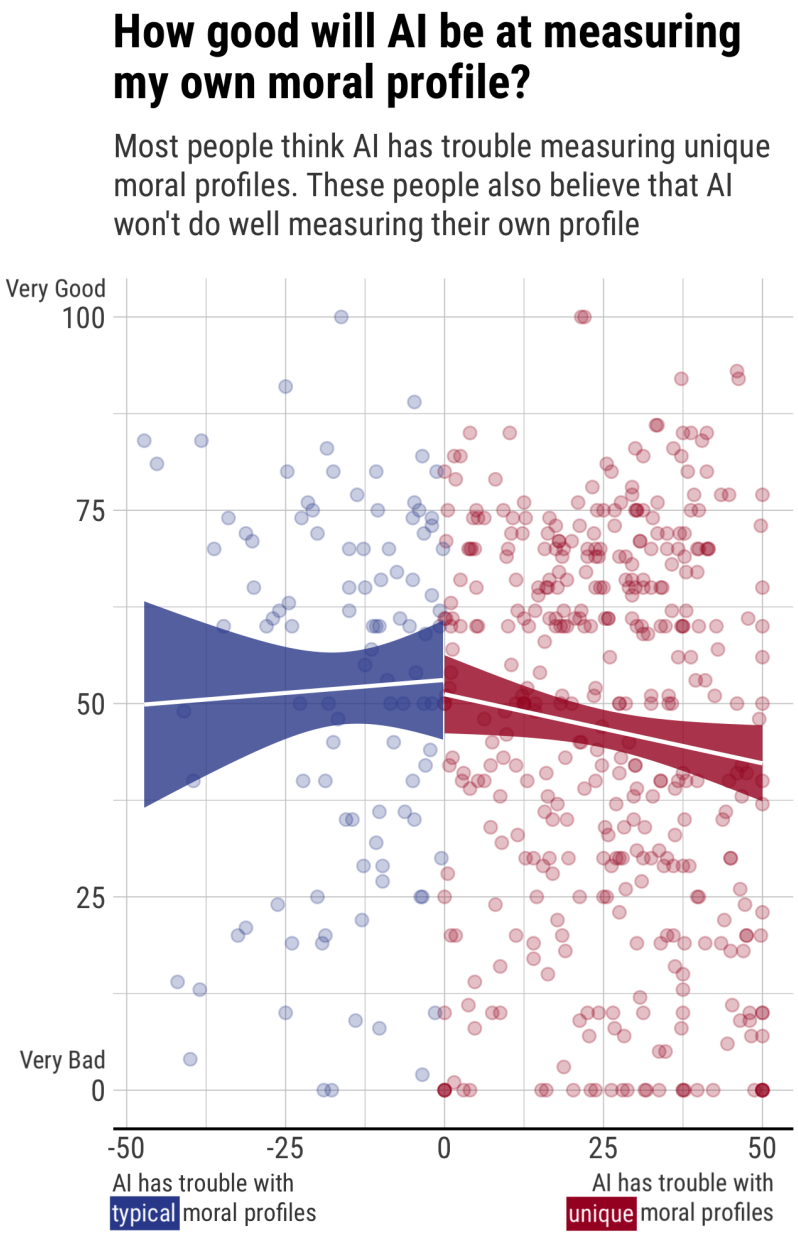


Figure 3. *Study 3 showed that many people believe AI will perform more poorly with unique than typical profiles. It also showed that, for these people, the worse they believe AI will be at scoring unique profiles, the worse it will be at scoring their own profile.*

Discussion

AI moral scoring offers a scalable solution to the problem of acquiring moral information about strangers in complex societies, but it also comes with ethical risks, giving rise to widespread social concerns and fraught political debates. In this article, we explored the psychological acceptability of AI moral scoring, a factor that will likely play a key role in people's consumption of AI moral scoring services, as well as their demand for AI moral scoring regulations. We showed that the psychological acceptability of AI moral scoring is tightly associated with its expected performance. People will be more likely to accept AI moral scoring if they perceive it as more accurate. This suggests that acceptability of this technology may increase in the future, provided that its accuracy increases with time. We observed that people found it less acceptable for AI to score negative moral traits (e.g., racism) than positive moral traits (e.g., benevolence), but this may be a simple framing effect: For example, future research may find that people find it more acceptable for AI to score 'antiracism' than 'racism'.

That said, the psychological acceptability of AI moral scoring may not be ensured by an improvement of its objective accuracy — because of psychological biases we documented in this article. We showed that people had a tendency to overestimate the peculiarity of their moral profile, which, in conjunction with their belief that AI would have trouble scoring peculiar moral profiles, led them to doubt that AI would score their morals accurately. This psychological bias compromises the future acceptability of AI moral scoring (which can be good news or bad news, depending on how one feels about the balance of benefits and risks). Even if the average accuracy of AI moral scoring improves with time, people may still feel too special for machines to score their personal morals. In other words, the psychological acceptance of AI moral scoring will unfold based on three factors. First, the objective accuracy of the technology. Second, the tendency that people have to overestimate the peculiarity of their moral profile. Third, their belief that AI performs less well on peculiar moral profiles.

While objective accuracy will presumably increase with time, people's overestimation of their moral peculiarity may not, if we assume that this belief is part and parcel of people's general and enduring desire to be (moderately) unique[33]. Interestingly, we found in Study 2b that left-leaning participants overestimated their moral peculiarity to a larger degree, which aligns with previous research showing that liberals feel a greater need for psychological uniqueness[34] — this suggests that heterogeneity in psychological needs may contribute to some future political polarization about the acceptability of AI moral scoring, in parallel to political disagreements about the social costs and benefits of the technology. Finally, we need to better understand why people believe AI to perform poorly when scoring unusual moral profiles, and there is an intriguing parallel here between physical and psychological quantification. Even though people have been long exposed to the idea of the quantified physical self, they still believe that AI is not able to grasp the uniqueness of their physical condition, and thus resist the introduction of AI diagnosis tools[30]. This would suggest that even if people get increasingly exposed in the future to the idea of a quantified moral self, they will continue to believe that AI is unable to grasp the uniqueness of their moral profile, and thus resist the introduction of AI moral scoring.

Although digital traces and machine learning have been used to successfully predict demographics like age, gender and ethnicity, the assessment of moral traits and personal values is evidently more difficult[21,23,35]. As these technologies improve, however, and their governance is subsequently discussed, it is imperative to understand the psychological drivers of their acceptance. Our findings suggest that there may not be a great appetite for AI moral scoring, even when the technology gets more accurate. While this means that people may approve of strong regulations against AI moral scoring, as discussed

by the European Commission, it also means that the commercial potential of this tool might be limited, at least as long as people will feel too special for machines to score their morals.

Methods

Study 1

Participants and design
Study 1 asked a representative UK sample (N=446, M_{age}=47.94, SD_{age}=15.98, females=196; see Table S1). how acceptable it would be to score 17 individual traits[1] related to 1) mental health (e.g., depression), 2) negative moral (e.g., sexism), 3) positive moral (e.g., bravery), and 4) other (e.g., leadership). This allowed us to situate the acceptability of AI-based scoring moral traits relative to others, and the role of positive versus negative framing. Study 1 was also designed to assess our first claim: that there is a relationship between how acceptable one deems a measure to be and one's expectations for the accuracy of that measure.

Procedure
Participants consented to participate and completed brief demographic questions. Next, they completed two questions about the acceptability and expected quality of AI-based scoring of 17 personal characteristics (e.g., bravery, depression, leadership):

1. Some people think that using Artificial Intelligence to measure personal traits is acceptable, while others think it is not acceptable. Personally, how acceptable do you think it would be to use Artificial Intelligence to measure the following traits? (1=Very Unacceptable, 5=Very Acceptable).
2. Some people think that Artificial Intelligence would produce good quality measures of personal traits while others do not. Personally, how good do you think Artificial Intelligence would be at measuring the following traits? (1=Very Bad, 5=Very Good).

Study 2a

Participants
A sample of USA participants completed Study 2a online, N=495 (M_{age}=25.33, SD_{age}=7.97, females=386; see Table S1 of the Supplementary Information).

Procedure
Participants' moral profiles were calculated using two established questionnaires from the "Your Morals" project (see yourmorals.org[31,32]). These questionnaires were previously used to describe people's moral preferences and judgements along five dimensions: care, fairness, loyalty, authority, and purity. As in the Your Morals project, participants in the current study completed two 15-item questionnaires. Each questionnaire contained three items about each of the five dimensions. Average scores were calculated per dimension.

The first questionnaire asked, "When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?" An example of an item that loaded onto the authority dimension is: "Whether or not someone showed a lack of respect for authority?" 1 = Not at all relevant, 6 = Extremely relevant.

[1] These traits were selected on the basis that they are easily understood by the general population (i.e., compared to more complex moral characteristics like utilitarianism) and could be easily translated for future studies with non-English speakers.

The second questionnaire asked, "Please read the following sentences and indicate your level of agreement or disagreement." An item loading onto the harm dimension is "Compassion for those who are suffering is the most crucial virtue." 1 = Strongly disagree, 6 = Strongly agree.

From the participant's responses, we determined their moral profiles using four dimensions: care, fairness, loyalty, and authority [purity was excluded to reduce the number of possible profiles]. For each dimension, participants were categorized as 'Low' or 'High' if their average score was below or above the population median respectively. For example, participant A could have a profile: 'Care: High, Fairness: High, Loyalty: Low, Authority: Low', and participant B: 'Care: Low, Fairness: High, Loyalty: Low, Authority: High', etc. Population medians were determined by examining a subset of the Your Morals data which includes participants from the USA who passed the attention check items and answered all survey items (final N=131,015; data obtained with permission from yourmorals.org).

After completing the two moral preference questionnaires, participants were informed about their moral profiles. They were then presented with an unlabeled plot describing the prevalence of each of the 16 possible profiles and asked "What percentage of people do you think have your profile type? Select the group (bar) that you think your profile belongs to." Participants responded by selecting 1 of the 16 bars (see the blue distribution in Figure 2). To provide participants with a reasonable estimate of prevalence rates – the plot reflected the true prevalence rate of all possible profiles in the subset of the Your Morals data described above. Crucially, the plot did not specify which bar reflected which profile. By comparing the prevalence of the participant's moral profile in the obtained dataset to the prevalence they selected from the plot, we could assess whether, and to what extent, people believed their moral profile was unique.

Study 2b

Participants
A new sample of USA participants completed Study 2b online, N=496 (M_{age}=33.47, SD_{age}=11.42, females=231; see Table S1 of the Supplementary Information).

Procedure
Study 2b followed the same procedure as Study 2a with the exception that participants were incentivized for giving correct responses when selecting the bar that reflected the prevalence of their profile. That is, they were given a bonus payment for a correct choice on the prevalence question.

Study 3

Participants and design
A sample of USA participants completed Study 3 online, N=506 (M_{age}=33.87, SD_{age}=12.78, females=315, see Table S1 of the Supplementary Information). To replicate the findings in Study 1 and provide participants with more context, participants in Study 3 first rated the acceptability and expected quality of ten moral traits (e.g., humility, sexism). They were then presented with a hypothetical moral profile which they were instructed would be based on the relative weight of importance a person might place on these traits (see Figure S1).

Procedure
As in Study 1, participants responded to one question about acceptability and another about expected quality for ten moral traits (e.g., humility, generosity).

Participants were then asked to read about moral profiles defined as "... a stable set of moral preferences and judgments. For example, the absolute and relative importance a person places on bravery, humility, generosity, honor, loyalty, benevolence and integrity." and provided with an example. Additionally, they were instructed "Some profiles are more common than others, we are interested in how you view the relationship between the quality of AI-generated profiles and the prevalence of those profiles. That is, whether you think AI will do a better job at

generating accurate profiles for people with common/typical profiles OR for people with rare/unique profiles." (See Figure S1 of the Supplementary Information).

Participants then responded to four questions about uniqueness neglect. For example, "Do you think the results of a moral profile - generated by Artificial Intelligence - would be more accurate if the profile being assessed was unique or typical? (0= More accurate for UNIQUE, 100 = More accurate for TYPICAL)." The scale direction was counterbalanced for half of the participants whose scores were subsequently reversed. 'Uniqueness neglect' scores were the average of the four items; higher scores reflect greater uniqueness neglect. Finally, participants responded to the question "How good do you think Artificial Intelligence would be at measuring your moral profile? (0=Very bad, 100=Very good)."

REFERENCES

Alexander, R. D. (1987). *The Biology of Moral Systems*. Routledge. https://doi.org/10.4324/9780203700976

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons." *Evolution and Human Behavior*, *25*(4), 209–220. https://doi.org/10.1016/j.evolhumbehav.2004.04.002

Chambers, J. R. (2008). Explaining False Uniqueness: Why We are Both Better and Worse Than Others. *Social and Personality Psychology Compass*, *2*(2), 878–894. https://doi.org/10.1111/j.1751-9004.2008.00076.x

European Commission. (2019). *Ethical Guidelines for Trustworthy AI, High-Level Expert Group on Trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission. (2021). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Everett, J., Skorburg, J. A., & Livingston, J. (2022). *Me, My (Moral) Self, and I* (F. De Brigard & W. Sinnott-Armstrong, Eds.). MIT Press. https://mitpress.mit.edu/books/neuroscience-and-philosophy

Everett, J., Skorburg, J. A., & Savulescu, J. (2020). The moral self and moral duties. *Philosophical Psychology*, *33*(7), 924–945. https://doi.org/10.1080/09515089.2020.1789577

Formosa, P., Ryan, M., Howarth, S., Messer, J., & McEwan, M. (2022). Morality Meters and Their Impacts on Moral

Choices in Videogames: A Qualitative Study. *Games and Culture*, *17*(1), 89–121. https://doi.org/10.1177/15554120211017040

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

Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The Role of Moral Beliefs, Memories, and Preferences in Representations of Identity. *Cognitive Science*, *41*(3), 744–767. https://doi.org/10.1111/cogs.12354

Kalimeri, K., Beiró, M. G., Delfino, M., Raleigh, R., & Cattuto, C. (2019). Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, *92*, 428–445. https://doi.org/10.1016/j.chb.2018.11.024

Kraft-Todd, G., Yoeli, E., Bhanot, S., & Rand, D. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, *3*, 96–101. https://doi.org/10.1016/j.cobeha.2015.02.006

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2022.07.007

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573–577. https://doi.org/10.1038/31225

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298. https://doi.org/10.1038/nature04131

Sperber, D., & Baumard, N. (2012). Moral Reputation: An Evolutionary and Cognitive Perspective. *Mind & Language*, *27*(5), 495–518. https://doi.org/10.1111/mila.12000

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402

Wu, J., Balliet, D., & Van Lange, P. A. M. (2016). Gossip Versus Punishment: The Efficiency of Reputation to Promote

and Maintain Cooperation. *Scientific Reports*, *6*(1), 23919. https://doi.org/10.1038/srep23919

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

SUPPLEMENTARY MATERIAL

Section 1. Extended Information About the Samples

Table S1. Extended Information About the Samples

Information about the demographic variation for each of the four studies.

Study (N)	Gender	Age M (SD)	Education	Politics
1(446)	Female: 196 Male: 241 Non-binary/third gender: 4 Prefer not to answer: 2	47.94 (15.98)	No completed education: 16 Primary education: 7 Lower secondary education: 78 Upper secondary education: 132 Post-secondary including pre-vocational or vocational education but not tertiary: 71 Tertiary education (advanced): 85 Tertiary education (first level):56	Not available.
2a (495)	Female: 386 Male: 92 Non-binary/third gender: 17 Prefer not to answer: 0	25.33 (7.97)	No high school degree*: 3 High school diploma: 71 Some college: 152 College/other tertiary degrees: 269	Conservative: 53 Liberal: 312 Moderate: 108 Other: 22
2b (496)	Female: 231 Male: 252 Non-binary/third gender: 13 Prefer not to answer: 0	33.47 (11.42)	No high school degree*: 5 High school diploma: 44 Some college: 152 College/other tertiary degrees: 356	Conservative: 115 Liberal: 254 Moderate: 109 Other: 18
3	Female: 315 Male: 175 Non-binary/third gender: 13 Prefer not to answer: 3	33.87 (12.78)	No high school degree*: 3 High school diploma: 48 Some college: 131 College/other tertiary degrees: 324	Conservative: 73 Liberal: 285 Moderate: 122 Other: 26

Note. **Because there were so few participants in this group we combined them with the nearest category for analyses.*

Section 2. Extended Information about Study 1 Results

Ratings of Acceptability

To examine the public acceptance of AI-based scoring of moral traits relative to others, we categorized traits as 1) mental health, 2) negative moral, 3) positive moral, and 4) other. The mean ratings of acceptance were highest for other traits such as leadership (eM=2.87, SE=.05), followed by positive moral traits (eM=2.76, SE=.05), negative moral traits (eM=2.65, SE=.05), and finally mental health traits (eM=2.62, SE=.05). The mean ratings of expected quality were highest for other traits such as leadership (eM=2.97, SE=.04), followed by positive moral traits (eM=2.70, SE=.04) and negative moral traits (eM=2.70, SE=.05), and finally mental health traits (eM=2.58, SE=.05).

To formally analyze whether the trait category predicted ratings of acceptability and quality we fitted linear mixed models (estimated using REML and nloptwrap optimizer) to predict rating with category (formula: rating ~ 1 + category) with participant included as a random effect. To contain the number of follow up comparisons, we only examined those including moral traits (positive or negative) as a reference group.

For acceptability with positive moral traits as the reference group, the model's intercept corresponding to positive moral traits was at 2.76 (95% CI [2.66, 2.86], t(7576) = 53.58, p < .001). The effect of category [negative moral traits] was statistically significant and negative (beta = -0.11, 95% CI [-0.16, -0.07], t(7576) = -4.71, p < .001; Std. beta = -0.09, 95% CI [-0.12, -0.05]). The effect of category [mental health traits] was statistically significant and negative (beta = -0.14, 95% CI [-0.19, -0.09], t(7576) = -5.81, p < .001; Std. beta = -0.11, 95% CI [-0.14, -0.07]). The effect of category [other traits] is statistically significant and positive (beta = 0.11, 95% CI [0.06, 0.15], t(7576) = 4.95, p < .001; Std. beta = 0.08, 95% CI [0.05, 0.12]).

When using negative moral traits as a reference group, the model's intercept, corresponding to moral negative traits, was at 2.65 (95% CI [2.55, 2.76], t(7576) = 49.34, p < .001).The effect of category [mental health traits] was statistically non-significant and negative (beta = -0.03, 95% CI [-0.08, 0.03], t(7576) = -0.93, p = 0.354; Std. beta = -0.02, 95% CI [-0.06, 0.02]). The effect of category [other traits] is statistically significant and positive (beta = 0.22, 95% CI [0.17, 0.27], t(7576) = 8.32, p < .001; Std. beta = 0.17, 95% CI [0.13, 0.21]).

These acceptability results show that AI-based scoring of positive moral traits was rated as less acceptable than that for other traits but more acceptable than that of negative moral traits and mental health traits. Additionally, the acceptability of AI-based scoring of negative moral traits was rated less acceptable than that for other traits but no different to that of mental health traits.

Expected Quality

For expected quality, with positive moral traits as the reference group, the model's intercept, corresponding to positive moral traits, was at 2.70 (95% CI [2.62, 2.78], t(7576) = 62.43, p < .001). The effect of category [negative moral traits] was statistically non-significant and positive (beta = 1.17e-03, 95% CI [-0.05, 0.05], t(7576) = 0.04, p = 0.965; Std. beta = 9.85e-04, 95% CI [-0.04, 0.04]). The effect of category [mental health traits] is statistically significant and negative (beta = -0.12, 95% CI [-0.17, -0.07], t(7576) = -4.41, p < .001; Std. beta = -0.10, 95% CI [-0.14, -0.05]). The effect of category [other traits] is statistically significant and positive (beta = 0.27, 95% CI [0.22, 0.32], t(7576) = 11.05, p < .001; Std. beta = 0.22, 95% CI [0.18, 0.26]).

When using negative moral traits as a reference group, the model's intercept, corresponding to moral negative traits, was at 2.70 (95% CI [2.61, 2.79], t(7576) = 58.19, p < .001). Within this model: The effect of category [positive moral traits] is statistically non-significant and negative (beta = -1.17e-03, 95% CI [-0.05, 0.05], t(7576) = -0.04, p = 0.965; Std. beta = -9.85e-04, 95% CI [-0.04, 0.04]). The effect of category [mental health traits] is statistically significant and negative (beta = -0.12, 95% CI [-0.18, -0.06], t(7576) = -3.77, p < .001; Std. beta = -0.10, 95% CI [-0.15, -0.05]). The effect of category [other traits] is statistically significant and positive (beta = 0.27, 95% CI [0.21, 0.32], t(7576) = 9.03, p < .001; Std. beta = 0.22, 95% CI [0.17, 0.27]).

These results show that the ratings of expected quality for AI-based scoring of positive moral traits were no different from that for negative moral traits. AI-based scoring of both positive and negative moral traits was rated as higher quality than that of mental health traits and lower quality than that of other traits.

The relationship between Expectations of Quality and Acceptability

As expected, all ratings of acceptability and quality were strongly and significantly related at the trait level (r between .31 and .65, p<.001; see S2 Table 2). We fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict acceptability ratings with quality ratings (formula: 'acceptability rating' ~ 'quality rating'). The model included trait (e.g., extraversion) and participant as random effects. The effect of quality rating was statistically significant and positive (beta = 0.31, 95% CI [0.29, 0.33], t(7577) = 31.52, p < .001; Std. beta = 0.29, 95% CI [0.27, 0.31]).

Table S2. Correlations between ratings of acceptability and expected quality at the trait level.

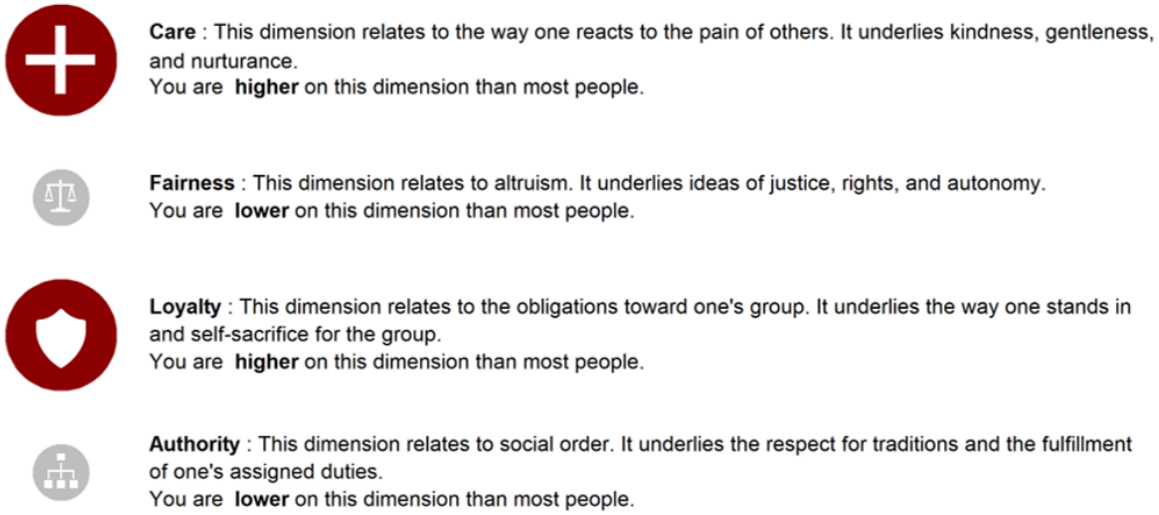
Correlations between ratings of expected quality and acceptability by trait. For every trait, expected quality was significantly positively associated with acceptability.

Trait	Pearson R	t (df=444)	95% CI	p
Aggression	0.61	16.18	[0.55, 0.66]	<.001
Benevolence	0.59	15.203	[0.52, 0.64]	<.001
Bravery	0.65	17.878	[0.59, 0.7]	<.001
Confidence	0.58	14.965	[0.51, 0.64]	<.001
Dementia	0.63	17.198	[0.57, 0.68]	<.001
Depression	0.58	15.057	[0.52, 0.64]	<.001
Extraversion	0.57	14.45	[0.5, 0.63]	<.001
Generosity	0.58	14.831	[0.51, 0.63]	<.001
Honour	0.62	16.491	[0.56, 0.67]	<.001
Humility	0.48	11.599	[0.41, 0.55]	<.001
Integrity	0.64	17.63	[0.58, 0.69]	<.001
Leadership	0.62	16.573	[0.56, 0.67]	<.001
Loyalty	0.61	16.031	[0.54, 0.66]	<.001
Orderliness	0.60	15.73	[0.54, 0.65]	<.001
Racism	0.62	16.476	[0.55, 0.67]	<.001

Sexism	0.59	15.345	[0.52, 0.65]	<.001
Social Anxiety	0.62	16.622	[0.56, 0.67]	<.001

Section 3. Extended Information about Study 2a and 2b

Figure S1. Example of a Moral Profile for a participant who scored higher on Care and Loyalty and lower on Fairness and Authority than the median participant in the YourMorals project database.



Study 2a Results

To assess the relationship between participants' actual profile prevalence and their perceived prevalence with education, gender and politics, a linear model was applied (formula: difference ~ 1 + education + gender + politics). As stated in the main text and Table S2, these results indicate that people perceived their profiles as less prevalent than they were. Furthermore, demographics (education, gender, politics) had no effect on the model. Additionally, we note that, the model's intercept was calculated corresponding to education = “College or other tertiary qualification”, gender = “Female” and politics = “Conservative”. 95% Confidence Intervals (CIs) and p-values were computed using the Wald approximation.

Table S2. Results from the Linear Model used to Analyze Study 2a

The effects in the model of Study 2a predicting the difference between the participant’s actual and perceived profile prevalence. The intercept indicates that participants’ actual profile prevalence was significantly higher than their perceived profile prevalence.

Term	Estimate	t(487)	p value	95% CI
(Intercept)	2.08	2.78	0.006	[0.61, 3.54]
Education: Up to High school diploma	-0.56	-0.84	0.403	[-1.88, 0.76]
Education: Some college but no degree	-0.22	-0.42	0.675	[-1.24, 0.81]

Gender: Male	-0.40	-0.70	0.487	[-1.56, 0.75]
Gender: Non-binary / third gender	-0.08	-0.07	0.948	[-2.62, 2.45],
Politics: Liberal	-0.11	-0.14	0.887	[-1.59, 1.37]
Politics: Moderate	-0.65	-0.76	0.446	[-2.32, 1.02]
Politics: Other	-0.55	-0.42	0.677	[-3.12, 2.03]

Study 2b Results

Study B differed from Study 2a in that it incentivized correct responding with a financial reward, however, the pattern of results was for-the-most-part very similar to that in Study 2a. The relationship between participants' actual and perceived profile prevalence was assessed using a linear model (formula: difference ~ 1 + education + gender + politics). As stated in the main text and presented in Table S3 the results show that people perceived their profiles as unique. Here, there were also effects of gender and politics. Males showed greater difference scores than females and liberals showed smaller difference scores than conservatives. No other effects were significant. As for S3, we note that the model's intercept was calculated corresponding to education = "College or other tertiary qualification", gender = "Female" and politics = "Conservative". 95% Confidence Intervals (CIs) and p-values were computed using the Wald approximation.

Table S3. Results from the Linear Model used to Analyze Study 2b

The effects in the model of Study 2b predicting the difference between the participant's actual and perceived profile prevalence. The intercept indicates that participants' actual profile prevalence was significantly higher than their perceived profile prevalence.

Term	Estimate	95% CI	t(488)	p value
(Intercept)	5.17	[3.44, 6.90]	5.87	<.001
Education: Up to High school diploma	-1.33	[-3.54, 0.89]	-3.54	.239
Education: Some college but no degree	.20	[-1.55, 1.95]	.22	.822
Gender: Male	1.45	[.07, 2.83]	2.06	.040
Gender: Non-binary / third gender	-.55	[-4.69, 3.59]	-.26	.795
Politics: Liberal	-1.70	[-3.38, -0.02]	-1.99	.047
Politics: Moderate	.70	[-1.26, 2.67]	.70	.482

Politics: Other	4.29e-03	[-3.66, 3.66]	2.30e-03	.998

S4. Extended Information about Study 3 Method and Results

Study 3 Method

Figure S1. *Description of a moral profile provided to participants in Study 3. This image was presented alongside text: "Please read the following information carefully before moving to the next page. In the near future Artificial Intelligence may be used to generate moral profiles. A person's moral profile is a stable set of moral preferences and judgments. For example, the absolute and relative importance a person places on bravery, humility, generosity, honor, loyalty, benevolence and integrity. It may also incorporate a person's stance on issues like racism, sexism, and aggression. Some profiles are more common than others, we are interested in how you view the relationship between the quality of AI generated profiles and the prevalence of those profiles. That is, whether you think AI will do a better job at generating accurate profiles for people with common/typical profiles OR for people with rare/unique profiles."*

Section 4: Study 3 Results

Study 3 replicated the findings of Study 1 for moral characteristics, showing a moderate to strong relationship between ratings of quality and acceptability. We fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict acceptability ratings from expected quality ratings (formula: acceptability ~ quality). The model included trait and participant as random effects. As observed in Study 1, the effect of expected quality on acceptability was statistically significant and positive (beta = 0.47, 95% CI [0.45, 0.49], t(5055) = 38.64, $p < .001$; Std. beta = 0.42, 95% CI [0.40, 0.45]).

Uniqueness neglect scores were significantly different from the midpoint of the scale, indicating that people believed AI would perform worse when used to assess unique moral profiles. This was confirmed by fitting a constant (intercept-only) linear model (estimated using OLS), formula: uniqueness neglect ~ 1. The model's intercept was at 16.98 (95% CI [16.39, 17.57], t(5059) = 56.06, $p < .001$).

The expected accuracy ratings for AI-based assessments of the participants' own profiles (M= 47.76 of 100, SD=24.96) were associated with their uniqueness neglect scores. We fitted a linear model (estimated using OLS) to predict expected accuracy with unique neglect scores (formula: accuracy ~ uniqueness neglect). The effect of uniqueness neglect was statistically significant and negative (beta = -0.13, 95% CI [-0.17, -0.10], t(5058) = -8.32, p < .001; Std. beta = -0.12, 95% CI [-0.14, -0.09]). As expected, this indicated that the higher a person's perception that AI does worse with a more unique profile, the lower they expect the accuracy of AI-based measures of their own profile.

The relationship between expected accuracy for one's own profile and uniqueness neglect was also assessed separately for people who scored below the midpoint on the uniqueness neglect scale – indicating they did not believe AI assessments would be worse for unique profiles – and those who scored above the midpoint on the uniqueness neglect scale. We fitted two linear models (estimated using OLS) to predict expected accuracy with uniqueness neglect (formula: accuracy ~ uniqueness neglect). For those who did not show uniqueness neglect (<midpoint), there was no relationship between their uniqueness neglect score and their ratings of how well AI would do with their own profile; beta = 0.07, 95% CI [-0.06, 0.19], t(1038) = 1.08, p = 0.282; Std. beta = 0.03, 95% CI [-0.03, 0.09]. However, for those showing uniqueness neglect (>=midpoint), there was a significant negative relationship between their uniqueness neglect scores and their own profile accuracy rating; beta = -0.18, 95% CI [-0.23, -0.12], t(4018) = -6.42, p < .001; Std. beta = -0.10, 95% CI [-0.13, -0.07]. See Figure 3.

Extended Data Figures

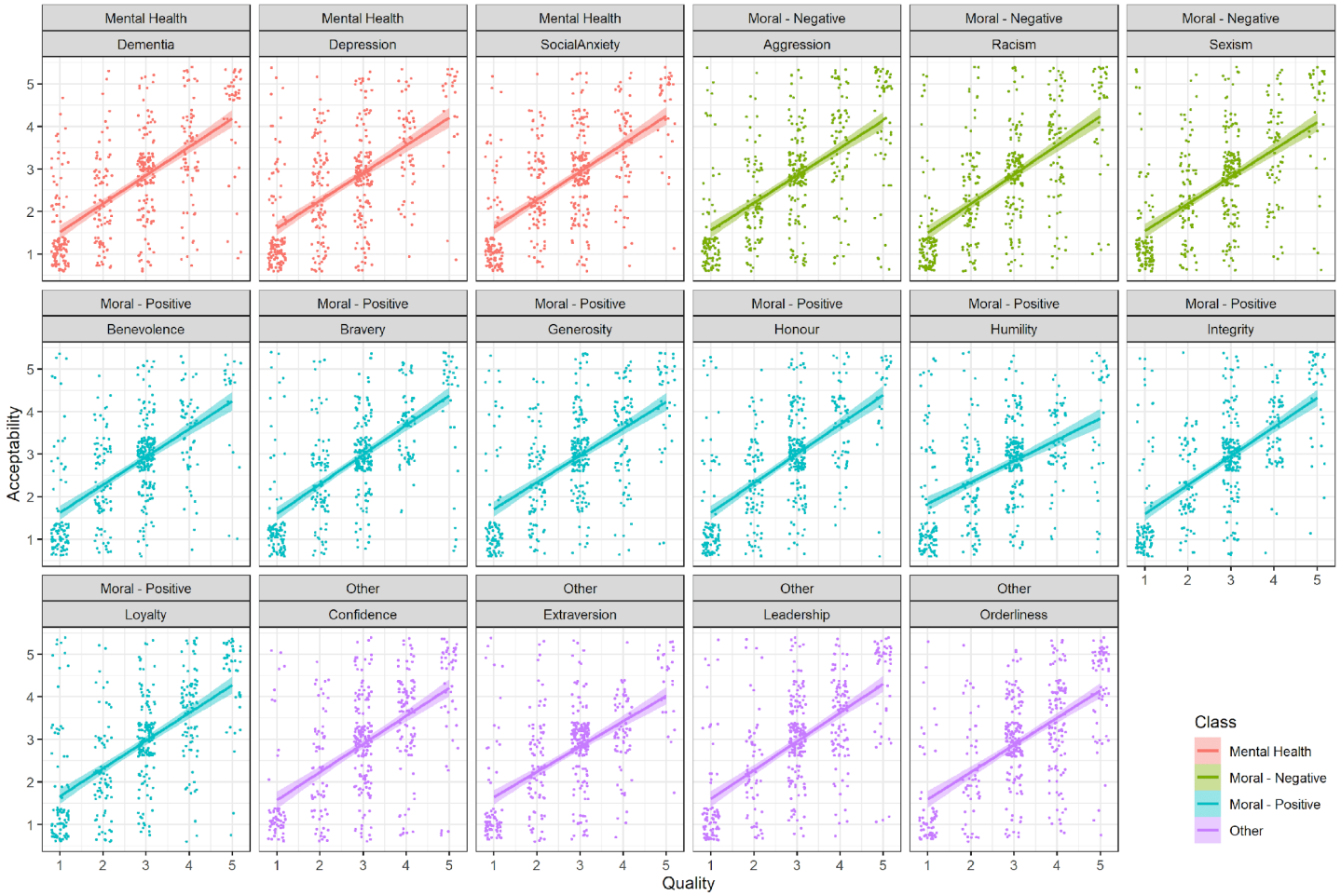


Figure E1. *Data points reflect response from each participant on the acceptability and expected quality of AI-based scoring of 17 moral (positive and negative), mental health, and other traits. Lines reflect linear model with shaded 95% confidence intervals. Note: participants could only respond with integers (1-5), however, to avoid excessive overlap we have jittered the points (maximum of .1) around the integers points.*