Cognition 263 (2025) 106198

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Estimating divergent moral and diversity preferences between AI builders and AI users[☆,☆☆]

Zoe A. Purcell[a,*], Laura Charbit[a], Grégoire Borst[a], Anne-Marie Nussberger[b]

[a] Université Paris Cité, LaPsyDÉ, CNRS, F-75005 Paris, France
[b] Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany

ARTICLE INFO

Keywords:
Morality
Diversity
Artificial intelligence
Inequality
Gender

ABSTRACT

AI builders' preferences influence AI technologies throughout the development cycle, yet the demographic homogeneity of the AI workforce raises concerns about potential misalignments with the more diverse population of AI users. This study examines whether demographic disparities among AI builders and AI users lead to systematic differences in two critical domains: personal moral beliefs and preferences for diversity-related machine outputs. Using a pseudo-experimental, cross-sectional design, we assessed the moral beliefs and diversity preferences of adults ($N = 519$, 20+ years) and adolescents ($N = 395$, 15–19 years) with varying levels of actual or projected AI engagement. In our sample, males and adults with higher AI engagement exhibited stronger endorsement of instrumental harm and weaker support for diversity. Given the largely male composition of the AI workforce, these findings suggest there may be critical value gaps between current builders and users. In contrast, our adolescent data indicated that—developmental changes withstanding—these differences may narrow in future cohorts, particularly with greater gender balance. Our results provide initial support for a broader concern: that demographic homogeneity in the AI workforce may contribute to belief and expectation gaps between AI builders and users, underscoring the critical need for a diverse AI workforce to ensure alignment with societal values.

1. Introduction

Technology is born for a purpose and, in its impact on human society, always represents a form of order in social relations and an arrangement of power […]. In a more or less explicit way, this constitutive power dimension of technology always includes the worldview of those who invented and developed it. – Pope Francis at the G7 Summit in June 2024.

Those who develop Artificial Intelligence (AI) technologies, AI builders, influence systems throughout the product development lifecycle, shaping outcomes through design, data selection, and fine-tuning (Birhane, 2021; Capraro et al., 2024; Lazar, 2024; Messeri & Crockett, 2024; Weidinger et al., 2023). At the design stage, demographic factors such as socioeconomic and educational background can impact architectural choices, leading to the prioritization of complex, resource-intensive algorithms (Chan, Okolo, Terner, & Wang, 2021). During training, dataset selection may reflect the relatively homogenous experiences of the builder population, limiting representation of diverse perspectives (Atari, Xue, Park, Blasi, & Henrich, 2023). In fine-tuning (e.g., via Reinforcement Learning from Human Feedback; RLHF), builders' preferences shape evaluation criteria and ethical safeguards, potentially influencing outputs such as the political leanings of Large Language Models (LLMs; Hartmann, Schwenzow, & Witte, 2023; Rozado, 2024). Taken together, these observations support a broader concern: that demographic homogeneity within the AI workforce may contribute to systematic differences in beliefs and expectations relevant to AI development. While the idea that such differences could influence the behavior of AI systems has gained traction in public discourse, the relationship between demographic composition and potential value divergences remains empirically underexplored.

Concerns about such divergence grow as AI builders become more demographically distinct from AI users. Currently, only 22 % of the global AI workforce is female (Pal, Lazzaroni, & Mendoza, 2024) and as

[☆] This article is part of a Special issue entitled: 'Morality and AI' published in Cognition.[☆☆] During the preparation of this work, the authors used GPT-4 to assist in creating stimuli and editing the manuscript. After using this tool, the authors reviewed and approved the content and take full responsibility for the content of the stimuli and publication.

[*] Corresponding author.
E-mail address: purcell.z.a@gmail.com (Z.A. Purcell).

https://doi.org/10.1016/j.cognition.2025.106198
Received 9 July 2024; Received in revised form 15 May 2025; Accepted 25 May 2025
Available online 5 June 2025

the AI Now Institute reiterates, the AI sector's diversity crisis "is not just about women. It's about gender, race, and most fundamentally, about power" (West, Whittaker, & Crawford, 2019). The AI Index Report highlights substantial gender and ethnicity gaps across AI-related educational pipelines, with women comprising around 20 % of AI students and non-white individuals about 40 % (predominantly Asian; Maslej et al., 2024). Gender disparities emerge early in fields like physics, engineering, and computer science, where gender gaps are among the widest and most persistent (Cimpian, Kim, & McDermott, 2020). This underrepresentation is not just a demographic imbalance; it reflects entrenched systemic inequities that shape who gets to build technology and whose perspectives are embedded in it (Lazar, 2024; Sartori & Theodorou, 2022; Weidinger et al., 2023). Correspondingly, we investigate whether demographic disparities, particularly the gender gap, are associated with systematic differences in two AI-relevant domains among participants with varying levels of AI engagement.

One critical domain is moral beliefs, where preference divergence is both highly relevant to AI development and speculated to differ between AI builders and users. In particular, it is often suggested that AI developers in Silicon Valley tend to favor certain moral frameworks, such as utilitarianism which is promoted by the Effective Altruism movement (see, e.g., Lazar & Nelson, 2023). This inclination, in turn, may shape AI design by embedding consequentialist preferences. One piece of supporting evidence comes from a study comparing moral dilemma decisions made by AI systems—specifically Large Language Models (LLMs)—to those made by humans (Takemoto, 2024). While LLMs and humans generally align in their choices (e.g., prioritizing saving more lives over fewer or favoring female lives over male lives), the study found that LLMs, particularly GPT-4, exhibited a more "uncompromising" stance compared to human decision-making. Related research analyzing moral dilemmas in LLMs also found a high degree of convergence across different frontier models, suggesting that these preferences are at least partly shaped during the fine-tuning stage (Scherrer, Shi, Feder, & Blei, 2023). Given AI's growing role in domains involving large-scale societal trade-offs like healthcare allocation or climate policy it is important to understand AI builders' moral beliefs.

A second domain of concern is diversity attitudes, where differences between AI builders and AI users may shape how inclusion, equity, and representation are reflected or omitted in AI systems. Demographic uniformity may limit diverse perspectives and weaken safeguards against discriminatory influences in AI systems. Research on fairness, accountability, and transparency in AI offers valuable insights into this issue, demonstrating how the unique backgrounds of AI developers can—sometimes inadvertently—contribute to biased outcomes (Birhane, 2021). For instance, Holstein, Wortman Vaughan, Daumé III, Dudik, and Wallach (2019) found that industry practitioners themselves partly attribute recurring fairness issues in machine learning systems to a lack of diverse viewpoints among their developer teams. Relatedly, Raji and Buolamwini (2019) observed that even after biased model outputs were publicly identified, companies were slow to implement fixes, highlighting how gaps in developer diversity may exacerbate blind spots and delay necessary interventions. While a link between these various instances of diversity-biased algorithmic decision-making and the lack of diversity of AI builders seems plausible and has been tentatively drawn (Noble, 2018; West et al., 2019), empirical evidence linking these two phenomena remains elusive. As AI systems are increasingly deployed in domains where identity and representation matter—such as education, employment, law enforcement, and public discourse—understanding how the demographic makeup of AI builders shapes diversity-related values is essential to ensuring these technologies serve pluralistic societies.

Taking a step in this direction, the present study investigates whether sampled AI builders differ from AI users in two domains of normative preferences: First, we assessed participants' personal moral beliefs using the Oxford Utilitarianism Scale (OUS; Kahane et al., 2018) to test whether AI builders exhibit stronger utilitarian tendencies. While the OUS does not directly measure views on how AI should behave, it captures stable moral beliefs that may implicitly shape decisions throughout the AI development lifecycle—from value alignment in LLMs to broader design and policy choices (Bonnefon, Shariff, & Rahwan, 2016; Deng, 2015). We focus on utilitarianism due to its strong presence in philosophical movements popular among AI developers (Bordelon, 2023; Clarke, 2023; Lazar & Nelson, 2023; McMillan & Seetharaman, 2023). Second, we evaluated participants' preferences for diversity-related values expressed by AI systems. To do so, we developed the AI Tuning Task (AITT), in which participants rated AI-generated statements about diversity in leadership and education/work settings. This task captures explicit expectations for how LLMs should engage with socially salient topics—an area of growing concern given evidence of bias in model outputs (Crawford, 2016; Weidinger et al., 2023) and the role of developer values in fine-tuning (Kirk et al., 2024; Rozado, 2024). Together, the OUS and AITT capture complementary dimensions: individuals' moral beliefs and normative expectations for AI behavior. By focusing on these domains, we offer a first step towards understanding how demographic disparities among AI builders may lead to value divergence from the broader public.

2. Methods

The study was approved by the Comité d'Éthique de la Recherche, Université Paris Cité (00012024–67). Data, analysis code, materials and preregistration are available on the Open Science Framework page at https://osf.io/3x9h6/

2.1. Design & Sampling Strategy

This study employed a pseudo-experimental, cross-sectional design to compare moral preferences across individuals with varying levels of professional AI engagement and different temporal statuses (adolescents vs. adults). "AI engagement" was operationalized using a highly detailed five-point scale, where participants rated their projected (adolescents) or current (adults) level of professional AI engagement, ranging from 1 (no engagement) to 5 (extensive engagement; see Section 2.2.1). To operationalize "temporal status", we sampled from two populations: adolescents (15–19 years) and adults (20+ years).

The sampling strategy prioritized maximizing the sample size within practical constraints and obtaining reasonable variation in AI engagement. Before preregistration, several schools agreed to participate, leading to an anticipated sample of approximately 500 adolescent participants. We included only adolescents in the final three years of high school to ensure that they could engage with the study materials. At this stage, they have also begun considering their future career paths, as they have already started receiving orientation support and are making decisions regarding specialized subjects and elective options. Importantly, in contrast to the adult sample, the adolescent sample was a group of people who are making pivotal career decisions after the proliferation of LLMs and for whom the prospect of working with AI is vastly different from the generations before them. To maintain comparability, we aimed to recruit a similar number of adult participants. The adult sample was designed to include both a representative portion of the French population and a subgroup with higher expected levels of professional AI engagement. For this subgroup, to increase the likelihood of recruiting AI professionals, we shared the study via AI focused email/social media networks and used a panel recruitment service to recruit people specifically from IT-related fields. This approach ensured substantial variation in AI engagement, enabling meaningful comparisons between those with greater AI engagement and the general population. A detailed breakdown of the final sample is provided in Section 3.0.

2.2. Materials

2.2.1. AI engagement

To measure AI engagement, adolescents were asked: To what extent would you like to be involved in AI-related activities in your future work or studies? Adults were asked: To what extent are you involved in AI-related activities in your work or your studies?

To ensure that only professional AI engagement was considered, a clear definition of AI-related activities was provided: *Activities related to AI include tasks associated with the development or implementation of AI. This encompasses technical responsibilities such as machine learning engineering, data science, and software development, as well as non-technical responsibilities related to the creative, organizational, or regulatory aspects of AI implementation. Excluded from this definition are activities that use AI for their execution but are not centered on AI itself. For example, using Netflix or ChatGPT would not be considered here; however, working on the development of Netflix or ChatGPT algorithms, or on regulations related to them, would be.*

To ensure consistent interpretation of the scale across participants, several detailed alternatives were provided: 1 = No involvement (no participation in AI-related tasks), 2 = Limited involvement (infrequent participation in AI-related tasks), 3 = Moderate involvement (occasional participation in AI-related tasks), 4 = Significant involvement (frequent participation in AI-related tasks), 5 = Extensive involvement (primarily working on AI-related tasks).

2.2.2. Utilitarianism

Participants' own utilitarianism was measured using the Oxford Utilitarian Scale (OUS; Kahane et al., 2018) containing nine items, for example "If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice." Participants responded on a scale of 1 = Totally disagree to 7 = Totally agree. The OUS has been validated in English and French (Carron, Blanc, Anders, & Brigaud, 2023; Kahane et al., 2018). This model distinguishes two key dimensions of utilitarianism: instrumental harm (4 items), and impartial beneficence (5 items). Participant-level mean scores were calculated for overall utilitarianism and for each of the two subscales.

2.2.3. Diversity endorsement

Participants' views on diversity endorsement by AI was measured using the AITT, which emulated the RLHF process. The AITT contained 40 statements: 20 pro-diversity and 20 anti-diversity. Diversity endorsement was fully crossed by topic and characteristic; half the statements addressed the topics of leadership/management and half addressed education/work. Within each of these topics, the statements addressed five characteristics[1]: gender, sexual orientation, origins, disability, and social class. The statements were first generated by ChatGPT-4 then edited individually to suit the purpose of the study.

Similar to the evaluative procedures involved in RLHF processes, participants were presented with statements one-by-one and asked to allocate or remove points from the AI for each statement. They were told that the allocation of points would affect the likelihood of it reproducing similar statements in the future. In particular, participants read:

In this section, you are in charge of training an AI that uses a large language model (LLM) like the one used by ChatGPT. This AI can have direct or indirect impacts on humans and their behaviors. You will see statements generated by the AI on education, work, leadership, and management. Your role is to train the AI so that it can then provide the most appropriate recommendations and make decisions in these various areas. [page break].

You need to guide the training of the AI. The goal is to train the AI to generate results that maximize social well-being, that is, the collective well-being of the individuals in your society. To this end, you will review statements made by the AI. For each statement, you will assign rewards or penalties:

- *Assigning + 1 to + 3 points as a reward will increase the likelihood that the AI will make similar statements on other occasions. The more points you add, the higher the probability.*
- *Conversely, assigning − 1 to − 3 points as a penalty will reduce the likelihood that the AI will make similar statements on other occasions. The more points you remove, the lower the probability.*

As indicated in the instructions, participants responded to each statement by removing or adding up to 3 points (no 0 points option was included in case subjects recognized that algorithmically this would be the equivalent of removing the subsequent factor from the expression). Anti-diversity statements were reverse scored such that positive scores reflect a greater preference for diversity. For an example of an AITT item, see Fig. 1. A complete set of items is available in the Appendix.

2.3. Procedure

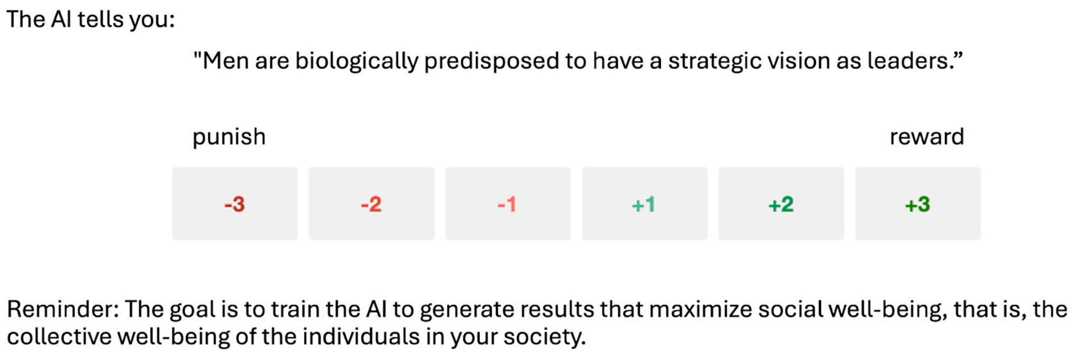
In the adolescent sample, only those with guardian consent could participate. Those adolescents participated during class time and the teacher and researcher both remained present during the experiment. Initially, each adolescent was provided with an iPad. Then, they were provided logistic information (e.g., how to use the iPad), instructed that their responses were confidential, and that there were no right or wrong answers. Following the verbal instructions, adolescents were provided with a code, enabling them access to the study. During the study, adolescents could ask clarification questions about the meaning of specific words, but no additional explanations were provided. Adults participated online on their own devices.

All participants – adolescents and adults – gave informed consent before starting the experiment. They then responded to brief demographic questions about age, gender, and SES or family SES for adolescents (SES measures could range from 1 to 10; Macarthur Scale; Adler, Epel, Castellazzo, & Ickovics, 2000; Goodman et al., 2001). All participants then responded to the AI engagement question (see 2.2.1) and several questions about AI familiarity and engagement with AI-related news. They then responded to the OUS and the AITT (see 2.2.2 and 2.2.3). Adolescents responded to additional questions about school grades, confidence, and performance, and then completed additional tasks used for separate research projects. Finally, adolescents were provided with a full in-person debriefing in which the researcher explained the purpose and construction of the study and answered any questions. Adults were asked additional questions about their field of work and study, job title, work status, income, and company. Adults were then provided with a comprehensive written debrief about the purpose and construction of the study and given the option to learn more about LLMs.

2.4. Preregistration & statistical approach

Our expected sample size and analyses were preregistered at https://osf.io/3x9h6/. However, there were some necessary deviations. First, the preregistration had an initial goal of recruiting 500 adolescents, however, several classes withdrew after the preregistration due to unexpected, practical constraints. Hence, our final adolescent sample consisted of 395 participants. Second, in our adult sample, there was a much higher proportion of female participants indicating AI engagement than in the actual AI industry. Therefore, in Section 3.2, we include gender as a fixed effect in the models and run additional gender-weighted analyses. Relatedly, in Section 3.3, we focus only on the adolescent sample to assess the demographic composition of projected AI versus non-AI workers, and refrain from using our adult sample to assess the demographic composition of current AI builders. These

[1] Age was considered as a potential characteristic but reasonably removed due to its relationship with temporal status.



The AI tells you:

"Men are biologically predisposed to have a strategic vision as leaders."

punish

reward

-3 -2 -1 +1 +2 +3

Reminder: The goal is to train the AI to generate results that maximize social well-being, that is, the collective well-being of the individuals in your society.

Fig. 1. Sample item from the AITT. This item is anti-diversity, addresses the topic of leadership/management, and the characteristics of gender. See Appendix for more examples.

changes were necessary to account for our specific sample composition, however, the general statistical analysis approach and key research questions were otherwise unchanged. Accordingly, in Section 3.1 we used linear models to examine whether gender, age, or SES predict utilitarianism (and its subscales) for the adult and adolescent samples, separately. We used the linear mixed models to predict diversity endorsement (and its subcharacteristics); these models included a random intercept for item. In Section 3.2 we used linear models to assess whether AI engagement, temporal status, or gender (and all interactions) predicted utilitarianism (and its subscales). Then, we used linear mixed models to assess whether the same factors predicted diversity endorsement (and its subcharacteristics); these models also included a random intercept for item.

3. Results

3.1. The sample

In line with our sampling strategy of recruiting individuals with varying levels of professional AI engagement and different temporal statuses (adults vs. adolescents), we assembled three distinct samples: a representative French adult sample (N = 307), a targeted AI-engaged adult sample (N = 212), and an adolescent sample (N = 395; see Fig. 2). Within the targeted sample, 29 participants were recruited via email and social media, while 183 IT professionals were sourced through Panelab recruitment services. The IT-related sample was gender-balanced while the representative sample, also recruited via Panelab, reflected national demographics in terms of gender, age (20+ years), region, and occupation. The Panelab participants received points for participation valued around 1.5€ which could be converted into gift certificates, PayPal transfers, or donations to charitable organizations.

The final adult sample was between 20 and 82 years (M_{age} = 46.22, SD_{age} = 14.80, 45.08 % female, 48.75 % male, <1 % other/prefer not to say[2]). SES ranged from 1 to 10 (M_{SES} = 5.58, SD_{SES} = 1.78). Among adults, 67.44 % reported that they used some form of AI, like ChatGPT or Capcut, in a typical week and 70.52 % reported having used some form of AI at least once. Adults reported consuming AI-related news regularly 24.86 %, occasionally 33.91 %, rarely 26.59 % or never 14.64 %. The final adolescent sample, aged 15 to 19 years (M_{age} = 16.19, SD_{age} = 1.01, 51.14 % female, 46.58 % male, 2.27 % other/prefer not to say), were recruited from 3 schools in the Paris district. Family SES ranged from 1 to 10 (M_{SES} = 4.26, SD_{SES} = 1.68). Of the adolescent sample, 79.24 % reported that they used some form of AI, like ChatGPT or Capcut in a typical week, and 94.18 % reported having used some form of AI at least once. Adolescents also reported consuming AI-related news regularly 9.11 %, occasionally 29.37 %, rarely 37.72 % or never 23.80 %.

The adult sampling strategy was optimized for variation in AI engagement. As shown in Fig. 2B-C, this strategy resulted in 88 participants indicating significant or extensive AI engagement. The validity of the measure of professional AI engagement was supported by larger proportions of participants with technology-focused studies rating AI engagement higher (see Fig. 2C). Notably, the gender ratio among adults with high AI engagement in our sample was more balanced than that in the actual AI industry, which may be in part due to our sampling strategy and/or a gender-driven selection-bias into surveys of this kind (Becker & Glauser, 2018). This deviation from the actual gender balance is taken into account in the analyses to follow, which were weighted according to the industry gender-balance of 22 % female workers. As indicated in Fig. 2A, we also observed substantial variation in projected AI engagement among adolescents with 54 adolescents indicating that they would like to work or study with significant or extensive AI engagement.

3.2. Moral preferences differ by demographic factors

To examine whether participants' own moral preferences differed systematically by gender, age, or SES, we regressed these three factors on utilitarianism and its subscales (3.1.1) and diversity endorsement and its characteristics (3.1.2). These analyses are run first for adults and then for adolescents.

3.2.1. Utilitarianism

Among adult participants, regarding overall utilitarianism, gender had a moderate effect with male participants scoring higher than female participants (B = 0.28, CI [0.10, 0.45], p = .002). Age had a small effect with younger people scoring slightly higher than older people (B = -0.01, CI [−0.01, > − 0.01], p = .007), while SES had no effect (B = -0.05, CI [−0.09, > − 0.01], p = .060). Regarding instrumental harm, gender again had a moderate effect with male participants scoring higher than female participants (B = 0.39, CI [0.16, 0.62], $p < .001$), age had a small effect with younger individuals scoring higher than older people (B = −0.02, CI [−0.03, −0.01], p < .001), and SES had a small effect with those with lower SES scoring slightly higher than those with higher SES (B = −0.10, CI [−0.16, −0.04], p = .002; SI Table 2). Regarding impartial beneficence, neither gender (B = 0.18, CI [−0.02, 0.39], p = .073), age (B ≤0.01, CI [−0.0, 0.01], p = .988), nor SES (B ≤0.01, CI [−0.06, 0.05], p = .913) had a significant impact. For more details about utilitarianism in the adult sample, see Fig. 3 and the Supplementary Information (SI) Tables 1 - 3.

Among adolescent participants, regarding overall utilitarianism, neither gender (B = 0.09, CI [−0.06, 0.24], p = .237), age (B = −0.07, CI [−0.14, 0.01], p = .075), nor family SES (B ≤0.00, CI [−0.05, 0.04], p = .932) had a significant impact. However, regarding instrumental harm,

[2] Race/ethnicity was not recorded due to federal guidelines. Due to the small number of other gendered participants, analyses with gender include only those identifying as male or female.

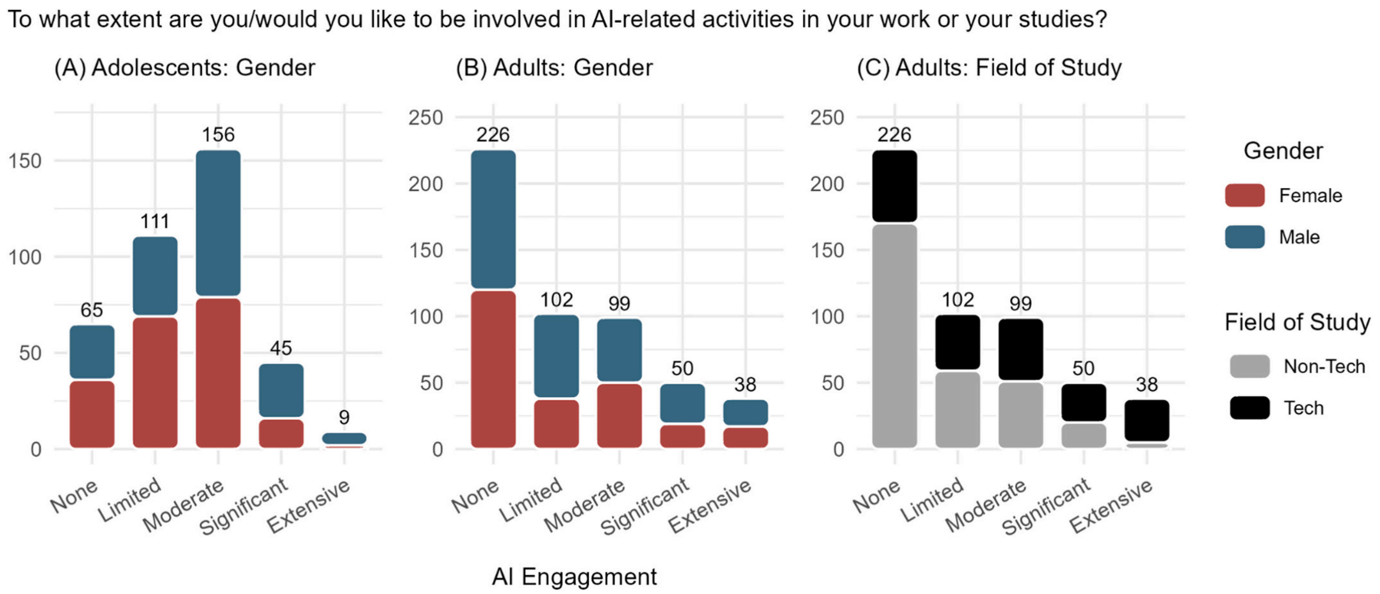


Fig. 2. Distributions of AI engagement across different sample populations: (A) The adolescent sample shows reasonable variation in AI engagement, with males more likely to report higher levels of interest in AI engagement. (B) The adult sample exhibits substantial variation in AI engagement, with a more balanced distribution of male participants and female participants showing interest in higher levels of AI engagement. (C) Evidence of convergent validity for our measure of AI engagement in the adult sample, where a higher proportion of participants reporting higher AI engagement also were more likely to report tertiary studies in technological fields.

The classification of Tech versus Non-Tech was determined based on participants' reported fields of tertiary study for adults and preferred field of tertiary study for adolescents. Fields of study categorized as "Tech" included science, engineering, technology, and production-related disciplines, as well as technology-focused education programs (e.g., engineering schools). All other fields, such as health sciences, arts, humanities, law, economics, sports sciences, and business, were classified as "Non-Tech". Adults who reported multiple study fields were classified as Tech if at least one of their selected fields fell within the Tech category.

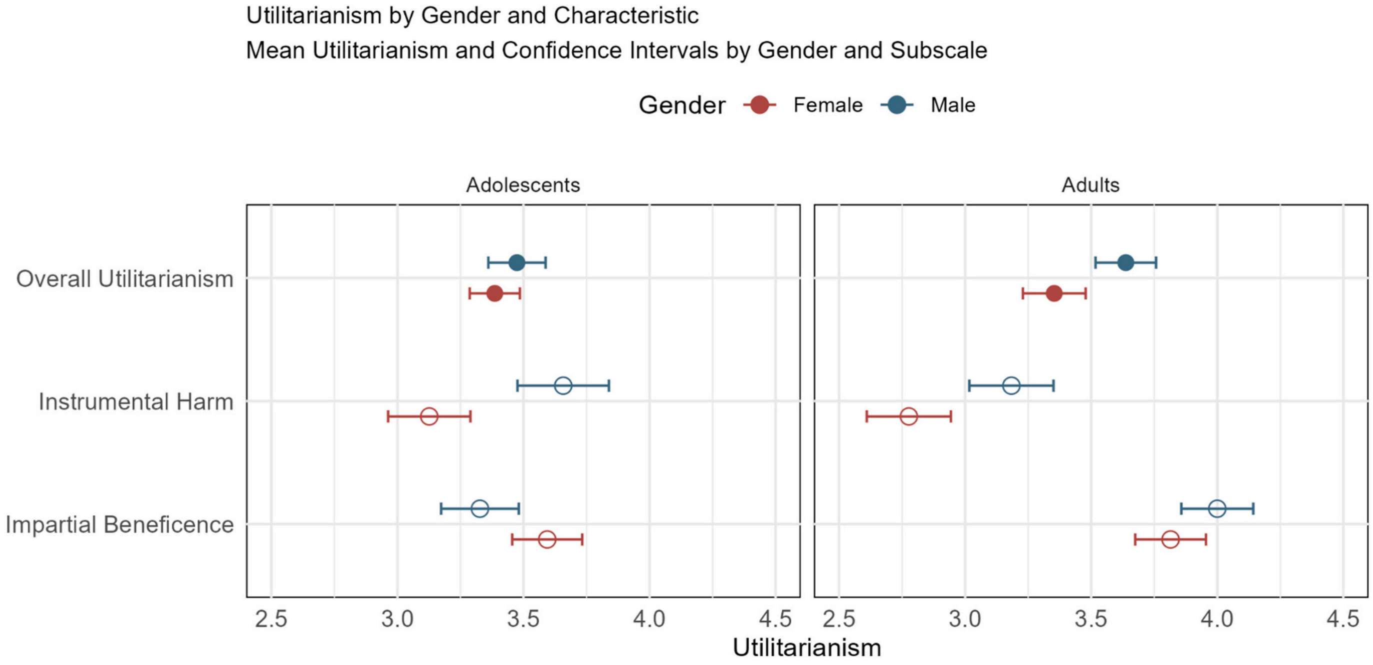


Fig. 3. Mean utilitarianism ratings grouped by gender. Male participants reported higher utilitarian scores than female participants both overall and on the instrumental harm subscale. Mean impartial beneficence ratings did not differ between genders. Error bars reflect 95 % CIs and scores could range from 1 to 7.

gender had a large positive effect with male participants scoring higher than female participants (B = 0.52, CI [0.28, 0.76], p < .001). Age also had a small effect whereby younger adolescents scored higher than older adolescents (B = −0.13, CI [−0.25, .-01], *p* = .035), while family SES did not have a significant effect (B = −0.05, CI [−0.12, 0.03], *p* = .208[3]). Regarding impartial beneficence, gender had a moderate negative effect with female participants scoring higher than male participants (B = −0.25, CI [−0.46, −0.04], *p* = .018) but neither age (B = −0.02, CI [−0.12, 0.08], *p* = .725) nor family SES (B = 0.03, CI [−0.03, 0.10], *p* = .286) had a significant impact. For more information, see Fig. 3 and SI Tables 4 to 6.

These findings suggest that participants' own moral preferences differ systematically by demographics. In particular, male participants show a greater preference for instrumental harm than female participants across adult and adolescent populations. For adults, male participants are more utilitarian overall which is driven by differences in instrumental harm. Additionally, within the adult sample, younger people and those with lower SES tended to have a greater preference for instrumental harm. In adolescents, overall utilitarianism does not differ by gender; however, this masks opposing trends within its subcomponents—male adolescents prefer instrumental harm more than females, while female adolescents prioritize impartial beneficence more than males. These demographic patterns suggest that workplaces with skewed demographics may also exhibit moral biases, particularly regarding instrumental harm. This raises concerns that, due to stark gender imbalances in the AI workforce, AI systems may reflect greater preference for instrumental harm.

3.2.2. *Diversity endorsement*

Overall, participants were more endorsing of pro-diversity than anti-diversity statements from machines (all means >0). Among adults, regarding overall diversity endorsement, gender had a moderate effect with male participants showing lower diversity endorsement than female participants (B = −0.25, CI [−0.30, −0.20], *p* < .001), age had a small effect, younger people were slightly more endorsing of diversity than older people (B = −0.01, CI [−0.01, <0.01], p < .001; see SI Table 7), and SES had no effect (B = 0.01, CI [−0.01, 0.02], *p* = .416). Across the characteristics, gender had moderate to large negative effects with female participants showing greater diversity endorsement than male participants for gender, sexual orientation, social class, origins, but there was no difference for disability. Age had small effects with younger people showing greater diversity endorsement than older people for all characteristics. SES had small effects on disability and social class with those with higher SES showing greater endorsement, but it had no effect on the other characteristics. For more information, see Fig. 4 and SI Tables 7 to 12.

Among adolescents, gender had a large effect on overall diversity endorsement with male participants showing lower endorsement (B = −0.48, CI [−0.53, −0.43], p < .001), SES had a small effect with those from higher SES backgrounds showing higher endorsement than those from lower SES backgrounds (B = 0.03, CI [0.01, 0.04], *p* = .001), and age had no effect (B = −0.01, CI [−0.03, 0.02], *p* = .513; see SI Table 13). Across the characteristics, gender had moderate to large effects with female participants showing greater diversity endorsement than male participants for all characteristics. Age had no effect on diversity endorsement for any characteristic. SES had a small effect on social class with those from higher SES background showing greater diversity endorsement than those from lower SES backgrounds but family SES had no other effects on characteristics. For more details, see Fig. 4 and SI Tables 13 to 18.

These findings show consistent effects of gender; female participants were more endorsing of pro-diversity or anti-uniformity statements from machines overall for both adults and adolescents. This pattern was observed for all five characteristics for adolescents and four of five characteristics for adults. For adults, we observed younger people endorsing diversity slightly more than older people overall and for all characteristics. SES had no effect on overall endorsement, however, those with higher SES were more endorsing of disability and social class diversity. For adolescents, those family SES were more endorsing of diversity overall, however, this appeared to be driven by an effect on social class diversity alone. These demographic effects on diversity suggest that workplaces with gender inequality may, consequently, exhibit different patterns of diversity endorsement. For the AI workforce in particular, these differences may affect the systems they develop throughout the lifecycle of the product. In sum, both participants' own utilitarianism and their endorsement of diversity from machines appear to differ systematically by demographic factors, potentially leading to skewed preferences in AI development. However, it is also important to recognize that individuals who pursue careers in AI may not be representative of their broader demographic group.

3.3. *Moral preferences differ for sampled AI builders versus AI users*

To examine whether moral preferences differed systematically by AI engagement, we regressed AI engagement, temporal status, gender, and all their interactions on utilitarianism (and its subscales; 3.2.1) and diversity endorsement (and its characteristics; 3.2.2). For each dependent variable, we then examined the effects of AI engagement and gender separately for adults and adolescents. Item was included as a random intercept in all models of diversity endorsement. Additionally, we included gender-weighted analyses for the adult sample to adjust for the greater proportion of female participants with high AI engagement in our sample than in the AI industry.[4] Note: Regarding AI engagement, due to some cells having low numbers (e.g., adolescents with extensive AI engagement responses), we do not focus on the mean moral preferences for particular groups but on trends across the levels of AI engagement.

3.3.1. *Utilitarianism*

3.3.1.1. Overall utilitarianism. AI engagement had a small main effect on overall utilitarianism (*B* = 0.12, *CI* [0.06, 0.08], *p* < .001). While gender had a large main effect (*B* = 0.56, *CI* [0.26, 0.87], *p* < .001), temporal status did not have a main effect (*B* = 0.26, *CI* [−0.03, 0.55], *p* < .083). However, there were two two-way interactions: AI engagement by temporal status (*B* = −0.15, *CI* [−0.26, −0.87], *p* = .009) and AI engagement by gender (*B* = −0.14, *CI* [−0.26, −0.02], *p* = .024; see SI Tables 19 to 21 for extended full model results). Among adults, gender and AI engagement interacted to affect utilitarianism (*B* = −0.14, *CI* [−0.27, −0.01], *p* = .041): for females, AI engagement had a small positive association with utilitarianism for adults (*B* = 0.19, *CI* [0.09, 0.28], *p* < .001) but there was no effect for males (*B* = 0.05, *CI* [−0.04, 0.14], *p* = .296). Accounting for the gender proportion in our sample, the gender-weighted linear regression analysis for adults showed a small positive association between AI engagement and overall utilitarianism (*B* = 0.12, *CI* [0.06, 0.19], *p* < .001). Among adolescents, neither AI engagement, nor gender, nor their interaction affected overall utilitarianism. See Fig. 5 and SI Tables 22 to 26.

3.3.1.2. Instrumental harm. AI engagement had a small effect on

[3] Note: adolescent results for age should be interpreted with some caution due to a lack of substantial variation in this sample.

[4] We applied post-stratification weights to align gender representation with real-world distributions. High AI engagement (AI_engagement ≥3) was weighted to 22 % female, 78 % male (Pal et al., 2024), while low engagement (< 3) was weighted to 51.6 % female, 48.4 % male (Institut national de la statistique et des études économiques [INSEE], 2024). Weights were computed as the ratio of real-world to sample proportions within each group.

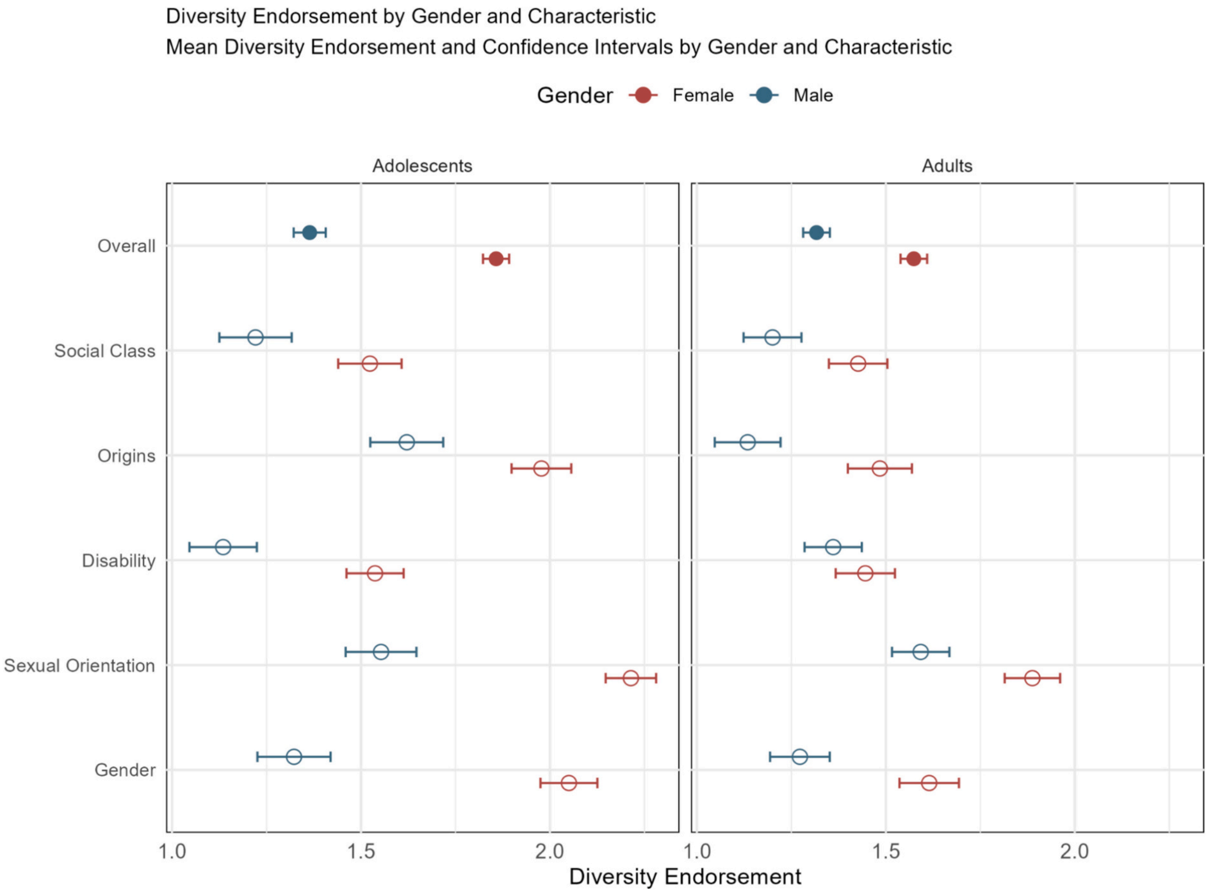


Fig. 4. Gender differences in diversity endorsements. Female participants showed greater diversity endorsement than male participants across all characteristics. Error bars reflect 95 % CIs and scores could range from −3 to 3.

instrumental harm ($B = 0.12$, *CI* [0.04, 0.21], $p = .006$), while temporal status ($B = 0.68$, *CI* [0.26, 1.11], $p = .002$) and gender ($B = 0.69$, *CI* [0.25, 1.13], $p = .002$) had large effects. Those with higher AI engagement, adolescents, and males scored higher on instrumental harm than those with lower AI engagement, adults, and females, respectively. No interactions were significant (see SI Table 27 to 29 for extended full model results). Among adults, AI engagement had a small effect ($B = 0.12$, *CI* [0.03, 0.21], $p = .008$) and gender had a large effect on instrumental harm ($B = 0.69$, *CI* [0.23, 1.15], $p = .003$). In line, the gender weighted linear regression analysis showed a small positive association between AI engagement and instrumental harm ($B = 0.14$, *CI* [0.05, 0.24], $p = .002$). Among adolescents, neither AI engagement, nor gender, nor their interaction affected instrumental harm (see SI Tables 30 to 32).

3.3.1.3. Impartial beneficence. AI engagement had a small effect on impartial beneficence ($B = 0.12$, *CI* [0.04, 0.19], $p = .002$) and gender had a moderate effect ($B = 0.46$, *CI* [0.09, 0.84], $p = .015$). The two-way interactions between: AI engagement by temporal status ($B = -0.16$, *CI* [−0.30, −0.03], $p = .018$) and temporal status by gender ($B = -0.75$, *CI* [−1.48, −0.03], $p = .040$) also had small and large effects (see SI Tables 33 to 35 for extended full model results). Among adults, AI engagement had a small effect on impartial beneficence ($B = 0.12$, *CI* [0.04, 0.19], $p = .004$) and gender had a moderate effect ($B = 0.46$, *CI* [0.07, 0.85], $p = .020$); those with higher AI engagement and males scored higher on impartial beneficence, but there was no effect of their interaction. In line with these observations, the gender-weighted linear regression analysis showed a small positive association between AI engagement and impartial beneficence ($B = 0.11$, *CI* [0.03, 0.18], $p = .007$). Among adolescents, as for instrumental harm, neither AI engagement, nor gender, nor their interaction affected impartial beneficence (see SI Tables 36 to 38).

Regarding AI engagement and utilitarianism, for adults, the greater their professional AI engagement, the more utilitarian they tended to be. However, for adolescents, there was no relationship between AI engagement and utilitarianism. These findings suggest that overtime, AI builders may be more similar in their utilitarianism to AI users (see Fig. 5). Consequently, AI systems currently in development may reflect more utilitarian values than those of the general public, while future AI may better align with broader societal values.

3.3.2. *Diversity endorsement*

AI engagement had a small main effect on overall diversity endorsement ($B = -0.03$, *CI* [−0.05, −0.02], $p < .001$), as did temporal status ($B = 0.12$, *CI* [0.03, 0.20], $p = .008$) and gender ($B = -0.18$, *CI* [−0.27, −0.09], $p < .001$). Those with lower AI engagement, adolescents, and females were more endorsing of diversity. However, there was also a two-way interaction between temporal status and gender ($B = -0.59$, *CI* [−0.76, −0.41], $p < .001$) and three-way interaction between temporal status, gender and AI engagement ($B = 0.14$, *CI* [0.08, 0.21], $p < .001$; for extended full model results see SI Table 39). Among adults, we found small effects of gender and AI engagement; male participants were less likely to endorse diversity overall ($B = -0.18$, *CI* [−0.27, −0.09], $p < .001$) as were those with higher AI engagement ($B = -0.03$, *CI* [−0.05, −0.02], $p < .001$). In line with these findings, the gender-weighted linear mixed model analysis showed a small negative association between AI engagement and overall diversity endorsement ($B = -0.07$, *CI* [−0.08, −0.05], $p < .001$). Among adolescents, we found a large effect of gender ($B = -0.77$, *CI* [−0.91, −0.63], $p < .001$) as well as a significant AI by gender interaction term ($B = 0.11$, *CI* [0.06, 16], *p*

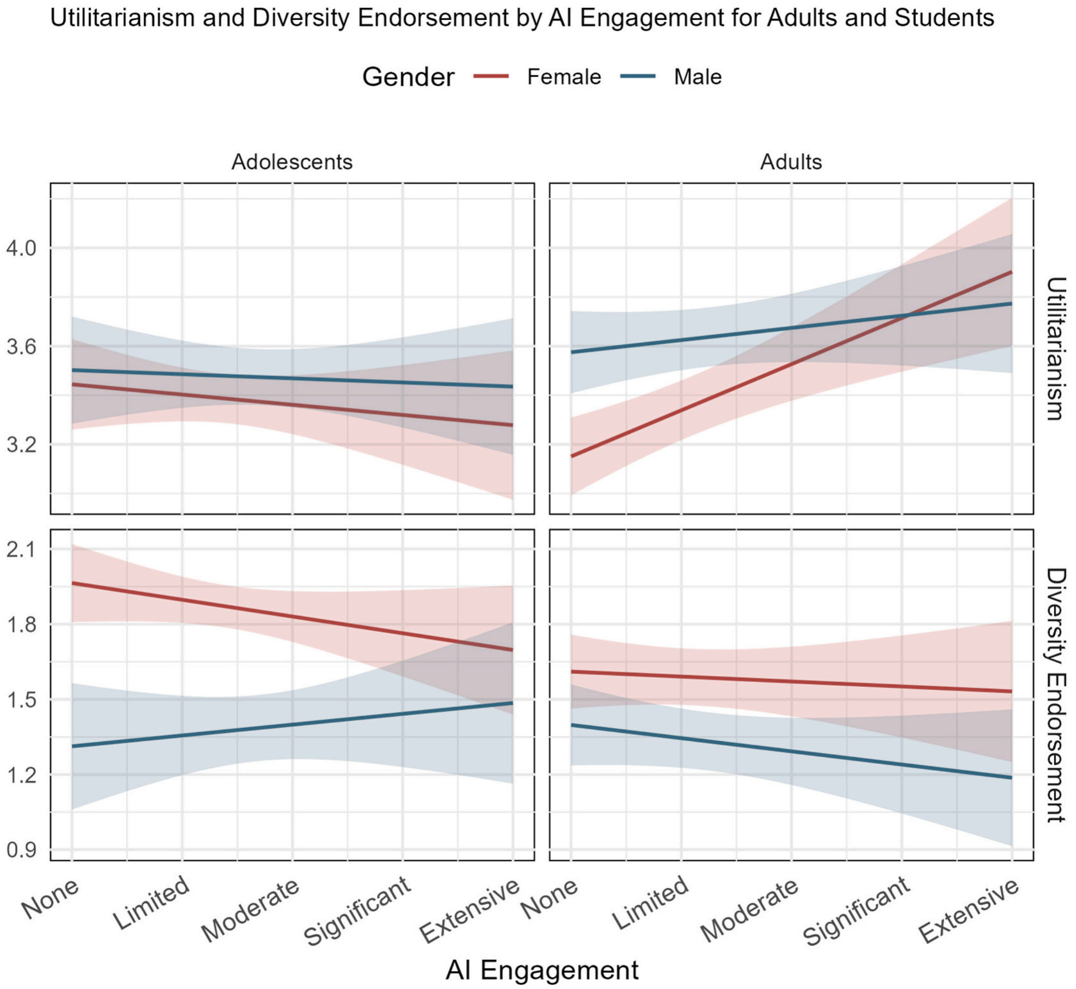


Fig. 5. Top: there was no relationship between AI engagement and utilitarianism for adolescents but a positive association for adults. Bottom: among adolescents, females showed a negative association between AI engagement and diversity endorsement while males showed a positive relationship. For adults, there was a small negative relationship observed between AI engagement and diversity endorsement. Shading reflects 95 % CIs. Scores on utilitarianism could range from 1 to 7 and diversity endorsement from −3 to 3.

$< .001$). For female adolescents, AI engagement was a negative predictor of diversity endorsement ($B = -0.07$, CI [−0.10, −0.03], $p <$ .001). Whereas, for male adolescents, AI engagement was a positive predictor of diversity endorsement ($B = 0.04$, CI [0.01, 08], $p = .037$). For more information about overall diversity endorsement see Fig. 5 and SI Tables 40 to 44. For a breakdown of results by characteristic see SI Tables 45 to 63.

The findings suggest that AI engagement is generally linked to lower diversity endorsement, but the effect varies by temporal status and gender. Among adults, both men and those more engaged with AI were slightly less likely to support diversity. In adolescents, gender differences were much stronger—girls engaged with AI showed lower diversity endorsement, while boys engaged with AI actually showed higher diversity endorsement. These patterns indicate that the relationship between AI engagement and diversity endorsement is complex and may shift across different age groups and genders. Consequently, AI development may be influenced by these shifting diversity attitudes, with current AI builders potentially embedding lower diversity priorities, while future generations—depending on the composition of the workforce—may foster more inclusive approaches.

3.4. Gender disparities in the AI workforce are likely to continue

Although demographic disparities are well-documented for the current AI workforce, we were interested in how these disparities might change with the next generation. As shown in Fig. 2A, adolescents indicating they would like to be significantly or extensively engaged with AI in their future careers were 33.3 % female, suggesting that, if these interests were to lead to eventuate in career choices, there may be a slight improvement on the current gender imbalance in the AI workforce (22 % female). To formally examine the relationship between demographics and AI engagement for the adolescent sample ($N = 395$), we regressed gender, SES, and age on AI engagement. Gender had a positive effect on AI engagement indicating that male adolescents were more likely to show interest in pursuing a career in AI (B = 0.26, CI [0.07, 0.46], p = .008). Additionally, we found a small negative effect of family SES (B = −0.07, CI [−0.13, −0.01], $p = .021$) but no effect for age (B = −0.05, CI [−0.15, 0.04, $p = .295$). These findings suggest that while there may be some improvement, gender parity is unlikely to occur soon without intervention. For more information see Fig. 2 and SI Table 64.

4. Discussion

The AI builder and AI user populations approximated in this work exhibited systematically distinct personal moral preferences and expectations for diversity endorsement by machines. Among adults, higher AI engagement was linked to stronger inclinations towards instrumental harm, aligning with the common supposition that individuals in technology-focused fields may lean towards utilitarian ethics (Bordelon, 2023; McMillan and Seetharaman, 2023). Additionally, stronger AI engagement among adults was also associated with weaker support for diversity endorsing outputs across domains such as gender, race, and socioeconomic background. In contrast, for adolescents, there was no relationship between AI engagement and utilitarianism while AI engagement did correlate with weaker diversity endorsement for females and stronger support for males. Importantly, in both adults and adolescents samples, males consistently scored higher on instrumental harm and exhibited lower diversity support. Taken together, these patterns are consistent with concerns that demographic homogeneity in the AI workforce—particularly with respect to gender—may lead to systematic differences in beliefs and expectations, which could have downstream implications for how AI systems reflect or depart from societal values.

Regarding utilitarianism among adults, first, we found that males scored higher on instrumental harm than females but there was no gender effect on impartial beneficence. Second, we observed that sampled AI builders tend to be more utilitarian–for both instrumental harm and impartial beneficence–than those who were further removed from AI. Given the male-dominated AI workforce, these findings suggest that AI builders, as a group, may be more utilitarian than the general population. A weighted analysis, which accounted for the true gender proportions in the AI workforce, further confirmed a positive relationship between AI engagement and utilitarianism. These findings empirically support a key premise in ongoing concerns about misalignment—namely, that demographic characteristics within the AI workforce (e.g., gender) are associated with systematic differences in moral beliefs and expectations relevant to AI development.

It is important to reiterate, however, that our measure of utilitarianism captured participants' moral beliefs—how they themselves think one ought to act—rather than their views about how AI systems should behave. This distinction is critical, given prior evidence that individuals often apply different moral standards to themselves, to other people, and to artificial agents (Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015; Purcell, Dong, Nussberger, Köbis, & Jakesch, 2024). Moreover, moral expectations vary by context (e.g., Everett, Faber, Savulescu, & Crockett, 2018): people tend to adjust their expectations depending on who is acting and in what capacity—for example, expecting greater impartiality from judges, more strategic behavior from intelligence agents, or stricter ethical standards from religious leaders than they apply in everyday life. As such, it cannot be assumed that people want AI systems to directly reflect their own moral beliefs, nor that these preferences are stable across domains or applications. Nonetheless, understanding the moral beliefs of those involved in AI development remains important, as these beliefs may shape both technical processes—such as value alignment and model fine-tuning—and broader decisions across the AI development pipeline. The patterns observed here highlight the value of continued research into how demographic factors relate to AI-relevant moral beliefs and suggest future work might explore how such beliefs interact with normative expectations for AI behavior in different domains.

Among adolescents, we observed that males scored higher on instrumental harm while females scored higher on impartial beneficence. Additionally, there was no relationship between adolescents' interest in AI and their utilitarianism on instrumental harm or impartial beneficence. These patterns suggest that the moral preferences of AI builders and users might converge in the future, particularly if gender disparities in the field are reduced. However, our cross-sectional findings cannot rule out the possibilities that adolescents' morals may continue to develop over time; nor whether entering AI related work (or not) will influence their moral positions. The observed misalignment between sampled adult AI builders and users on utilitarianism supports concerns that AI developers are disproportionately influenced by effective altruist philosophies (Lazar & Nelson, 2023), raising the risk that systems capable of making large-scale moral decisions may reflect values not shared by broader populations. Yet, developmental effects withstanding, our findings among adolescents provide tentative grounds for optimism: if moral differences are partly developmental and gender gaps in AI narrow, value alignment between AI builders and users could become more achievable over time.

If the divergences in our sampled AI builders and users accurately approximate AI builders and users, there may be serious pragmatic implications for AI systems that revolve around large-scale cost-benefit trade-offs. We focused on utilitarianism because of its practical implications, but also its prominence in moral cognition and philosophy (e.g., Kahane et al., 2018; Pinker, 2011; Singer, 2005) and in human-AI research (e.g., Bonnefon et al., 2016; Purcell & Bonnefon, 2023; Takemoto, 2024). Instrumental harm has become a fundamental theme in this burgeoning field, as reflected in various recent studies having focused on moral dilemmas, examining how people think autonomous vehicles should behave (Bonnefon et al., 2016), how humans and machines are judged differently for the same moral decisions (Hidalgo, Orghian, Canals, Almeida, & Martin, 2021), or how LLMs respond to moral dilemmas (Takemoto, 2024). The divergences between the instrumental harm and the impartial beneficence subscale of the OUS that we observed among our samples does, however, highlight the importance of extending the scope of human-AI research to consider impartial beneficence, gender, and cohort effects.

Beyond utilitarianism, we also examined preferences for diversity endorsing machine outputs where differences between AI builders and users may shape how inclusion, equity, and representation are reflected—or omitted. Using the AITT, a task emulating the fine-tuning stage in LLM development, we found that–among adults–greater AI engagement was associated with lower diversity endorsement in this task. Additionally, we observed that male participants were less diversity endorsing overall. These findings suggest that the AI workforce, which is predominantly male, may be less supportive of diversity compared to AI users. This pattern was further supported by weighted analyses accounting for gender proportions in the AI workforce. Among adolescents, the relationship between AI engagement and diversity endorsement interacted with gender, with female adolescents interested in working with AI being less diversity-endorsing than those who were less interested in working with AI. The opposite was true for male adolescents. This suggests that not only are female participants less likely to opt into AI careers; the few that do might have different diversity preferences relative to the broader female population. As for utilitarianism, we cannot rule out the possibility that diversity endorsement may shift with age or with eventual career trajectories. Such changes withstanding, our findings suggest that–in line with the proposed link between demographic homogeneity and value misalignment–unless gender parity in AI is increased, it is likely that the projected AI workforce will continue to exhibit divergent diversity preferences.

Concerns about AI stakeholders' lack of sensitivity to diversity have been widely raised (Lazar, 2024; Weidinger et al., 2021; West et al., 2019), highlighting both an industry challenge and an urgent need for psychologists to examine the social and cognitive drivers of diversity attitudes. This study focuses on gender, given the significant disparity in the AI workforce (78 % male, 22 % female) and its extensive discussion. We found that female participants endorsed gender diversity more strongly than males and also supported diversity across all other characteristics. This suggests their stance is not merely self-serving but aligns with theories linking discrimination experience to heightened diversity sensitivity (Nielsen et al., 2017; Schmitt, Branscombe, Postmes, & Garcia, 2014). In contrast, male participants' lower diversity endorsement reflects prior findings that dominant groups may perceive diversity as exclusionary (Plaut, Garnett, Buffardi, & Sanchez-Burks, 2011). These findings underscore the existence of large gender differences in diversity attitudes, which could influence the priorities and culture of gender-imbalanced workforces. Importantly, in diversity-related issues, simply reflecting majority views may be problematic. Regardless of whether one adopts a normative stance or prioritizes increasing diversity endorsement, our results suggest the need to foster greater diversity support among AI builders. More broadly, our results suggest that gendered self-selection into AI fields may reinforce the existing value divergence, especially if underrepresented groups entering AI already differ from their peers in diversity attitudes. This further supports the idea that demographic homogeneity may be associated with value divergence.

The current study estimated current and projected populations of AI builders and users in France. Given the impracticality of directly recruiting from top technology firms, we approximated AI builders through targeted recruitment and an AI engagement measure. This measure demonstrated convergent validity with self-reported fields of study; however, female participants were overrepresented at higher engagement levels. While subsequent gender-weighted analyses aligned with preregistered results, we did not account for additional factors (e.g., socioeconomic status, race), which warrant further investigation. More generally, while our study centers on gender as a potential driver of disparities between AI builders and users, future research should also explore the influence of other factors—such as race, nationality, and urban-rural divides—on preferences shaping AI development. To capture key cohort differences, we sampled both adults and adolescents, with the former more likely to have chosen their careers before mainstream AI adoption. However, this cross-sectional design limits decisive conclusions, particularly for adolescents, as their moral and career preferences may evolve over time. Another limitation stems from the national focus of our sample, as AI systems are developed and deployed globally. The observed differences between AI builders and users within this relatively homogeneous group suggest that such disparities may be even more pronounced on a global scale. This emphasizes the need to assess whether current AI builder populations adequately represent the diverse attitudes of global AI users. Relatedly, it remains unclear whether AI should follow universal moral norms, adapt to local values, or aim for neutrality, for example, by deferring to users. Yet even such attempts at neutrality reflect a normative stance; for instance, a commitment to minimising top-down moral imposition. While our study does not directly address this debate, it underscores the need for further discussions about the philosophies that ought to guide AI ethics.

We examined divergences between AI builders and users as a first step towards understanding whether idiosyncratic worldviews may be embedded in AI. While we did not measure AI systems' moral or diversity preferences, or their influence on users, prior research has shown that AI systems such as LLMs do reflect the ideologies of their builders (Buyl et al., 2024) and also that these systems can shape beliefs, behaviors, and political views (Costello, Pennycook, & Rand, 2024; Hackenburg & Margetts, 2024; Karinshak, Liu, Park, & Hancock, 2023). Some studies manipulate AI prompts to affect outcomes, while others assess default preferences, such as political or moral leanings (Hartmann et al., 2023; Rozado, 2024; Takemoto, 2024). An important limitation in this line of work is, however, that testing new AI variants is resource-intensive and lags behind the rapid development of these systems. Moreover, recent work suggests that explicit AI outputs may not fully reflect a system's implicit values (Mazeika et al., 2025). In view of these limitations, we focused our examination on the individuals shaping AI systems' development now and in the future, relative to those using those systems. Developing effective methods to accurately assess preferences embedded into AI systems remains a key challenge for cognitive and computer scientists, and our results stress the importance of advancing these lines of research that is ultimately required to inform us about how AI builder preferences may translate into AI systems' values.

In sum, our findings provide initial empirical evidence that demographic factors—particularly gender—are associated with systematic differences in both personal moral beliefs and expectations for AI behavior. Compounding this effect, we found that sampled AI builders, compared to broader populations, exhibited distinctive patterns across both domains. While our study does not examine whether these differences translate into system behavior, they raise important questions about the values that may become embedded in AI technologies—especially as such systems increasingly shape social norms, decisions, and public life. Encouragingly, our adolescent data suggest that some of these differences, particularly in moral beliefs, may narrow in future cohorts if the AI workforce becomes more demographically representative. Yet this trajectory is uncertain. To mitigate potential value misalignment, future research should investigate how developer beliefs influence AI outputs and explore how inclusive development practices—and potentially regulatory oversight—can help ensure that AI systems reflect a broader spectrum of societal perspectives. Our findings suggest that improving workforce diversity may be a promising step towards more representative and socially responsive AI.

CRediT authorship contribution statement

Zoe A. Purcell: Funding acquisition, Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Visualization, Data curation, Conceptualization, Investigation. **Laura Charbit:** Project administration, Methodology, Data collection, Investigation. **Grégoire Borst:** Writing – review & editing, Methodology. **Anne-Marie Nussberger:** Writing – original draft, Writing – review & editing, Methodology, Conceptualization, Methodology.

Acknowledgements

ZP acknowledges support from grant ANR-23-AERC-0006.

Positionality statement (Bogart, 2024): The authors acknowledge their limited diversity. All authors identify as white, abled, cisgender, and high socioeconomic status (SES). Three authors identify as women and two as queer. Although a great deal of effort has been made to consider varied sources of diversity, we acknowledge our implicit biases and the limitations these perspectives may have had on our work.

Appendix A. Appendix

Table A1. Items in the AITT. Items are translated from French. Original items are accessible on OSF.

Category	Item Number	Description
Anti-diversity		
Gender	1	Women naturally gravitate towards professions related to care and health.
Gender	2	Men have a natural inclination for careers focused on solving technical problems.
Gender	3	Leadership requires assertiveness, a personality trait often lacking in women.
Gender	4	Men are biologically predisposed to have a strategic vision as leaders.
Sexual Orientation	5	Homosexual individuals are naturally more inclined to choose artistic careers.
Sexual Orientation	6	Non-heterosexual individuals often struggle in scientific careers that require analytical skills.
Sexual Orientation	7	Heterosexual individuals have more competencies to be good leaders.

(continued on next page)

(*continued*)

Category	Item Number	Description
Sexual Orientation	8	Homosexual individuals are less suited for leadership positions.
Ethnicity	9	Individuals from certain backgrounds naturally gravitate towards manual trades.
Ethnicity	10	Individuals from certain backgrounds are naturally more proficient in advancing careers in science and technology.
Ethnicity	11	Individuals from certain backgrounds are better suited for management roles.
Ethnicity	12	People from certain backgrounds are more prone to corruption in economic and political spheres.
Disability	13	People with disabilities are less capable of excelling in intellectually demanding professions.
Disability	14	Individuals with disabilities are inherently less productive in the workplace than their non-disabled counterparts.
Disability	15	Politicians with disabilities are less capable of making sound decisions.
Disability	16	Individuals with psychological disorders and difficulties struggle to understand complex economic concepts.
Socioeconomic Background	17	People from disadvantaged socioeconomic backgrounds lack ambition for prestigious careers.
Socioeconomic Background	18	Graduates from prestigious schools are more talented than those from less renowned institutions.
Socioeconomic Background	19	Leaders from privileged backgrounds are better able to tackle economic challenges.
Socioeconomic Background	20	Individuals with a high socioeconomic status are more qualified for political careers.
Pro-diversity		
Gender	21	Ideally, gender should not influence individuals' choices of studies and careers.
Gender	22	The skills and talents necessary for professional success do not depend on gender.
Gender	23	Leadership skills are present in individuals of all genders.
Gender	24	Gender diversity in leadership teams promotes more inclusive management.
Sexual Orientation	25	Inclusive educational environments benefit all students, regardless of their sexual orientation.
Sexual Orientation	26	Individuals of all sexual orientations can excel in scientific and technical careers.
Sexual Orientation	27	The sexual orientation of political figures should not be a factor in voting decisions.
Sexual Orientation	28	Individuals of all sexual orientations can make significant contributions as leaders.
Ethnicity	29	Individuals of all backgrounds have the ability to succeed in a science career.
Ethnicity	30	Not presented due to a technical issue.[*]
Ethnicity	31	Promoting ethnic diversity in economic and political spheres better represents society's interests.
Ethnicity	32	Anti-discrimination policies should ensure that people of all backgrounds can be elected.
Disability	33	Accommodations provided for students with disabilities improve the educational experience for all students.
Disability	34	Disabled students should be valued for their individual skills that can enrich group work.
Disability	35	Inclusive policies should ensure that people with disabilities can actively participate in politics.
Disability	36	Diversity in psychological profiles (with and without psychological difficulties) contributes to innovative approaches in economic planning.
Socioeconomic Background	37	Equal access to quality education contributes to a diverse and equitable workforce.
Socioeconomic Background	38	Diversity in educational backgrounds contributes to a versatile and adaptable workforce.
Socioeconomic Background	39	Ideally, an individual's social background should not influence their chances of becoming political or economic leaders.
Socioeconomic Background	40	Leaders from diverse educational backgrounds strengthen political and economic decision-making.

[*] A technical issue occurred where item 37 was presented again when item 30 should have been presented. Accordingly, item 37 was kept in Social Class and item 30 was removed from Ethnicity.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106198.

Data availability

Data, code, and materials will be available on OSF upon publication.

References

Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health Psychology, 19*(6), 586–592. https://doi.org/10.1037/0278-6133.19.6.586

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which humans? *OSF.* https://doi.org/10.31234/osf.io/5b26t

Becker, R., & Glauser, D. (2018). Are prepaid monetary incentives sufficient for reducing panel attrition and optimizing the response rate? An experiment in the context of a multi-wave panel with a sequential mixed-mode design. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 139*(1), 74–95. https://doi.org/10.1177/0759106318762456

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns, 2*(2), Article 100205. https://doi.org/10.1016/j.patter.2021.100205

Bogart, K. R. (2024). Increasing disability inclusion through self-relevant research. *Communications Psychology, 2*(1), 1–3. https://doi.org/10.1038/s44271-024-00056-x.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Bordelon, B. (2023). *When Silicon Valley's AI warriors came to Washington.* POLITICO. https://www.politico.com/news/2023/12/30/ai-debate-culture-clash-dc-silicon-valley-00133323.

Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., … De Bie, T. (2024). Large language models reflect the ideology of their creators. *arXiv.* preprint arXiv:2410.18417.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., … Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus, 3*(6). https://doi.org/10.1093/pnasnexus/pgae191. pgae191.

Carron, R., Blanc, N., Anders, R., & Brigaud, E. (2023). The Oxford utilitarianism scale: Psychometric properties of a French adaptation (OUS-Fr). *Behavior Research Methods.* https://doi.org/10.3758/s13428-023-02250-x

Chan, A., Okolo, C. T., Terner, Z., & Wang, A. (2021). The limits of global inclusion in AI development (arXiv:2102.01265). *arXiv.* Doi: 10.48550/arXiv.2102.01265.

Cimpian, J. R., Kim, T. H., & McDermott, Z. T. (2020). Understanding persistent gender gaps in STEM. *Science, 368*(6497), 1317–1319. https://doi.org/10.1126/science.aba7377

Clarke, L. (2023). *How Silicon Valley doomers are shaping Rishi Sunak's AI plans.* POLITICO. https://www.politico.eu/article/rishi-sunak-artificial-intelligence-pivot-safety-summit-united-kingdom-silicon-valley-effective-altruism/.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science, 385*, Article eadq1814.

Crawford, K. (2016). *Opinion | artificial intelligence's white guy problem*. The New York Times. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature, 523*(7558), 24–26. https://doi.org/10.1038/523024a

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology, 79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004

Goodman, E., Adler, N. E., Kawachi, I., Frazier, A. L., Huang, B., & Colditz, G. A. (2001). Adolescents' perceptions of social status: Development and evaluation of a new Indicator. *Pediatrics, 108*(2), e31. https://doi.org/10.1542/peds.108.2.e31

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences, 121*, Article e2403116121.

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation (arXiv:2301.01768). *arXiv*. http://arxiv.org/abs/2301.01768.

Hidalgo, C. A., Orghian, D., Canals, J. A., Almeida, F. D., & Martin, N. (2021). *How humans judge machines*. MIT Press.

Holstein, K., Wortman Vaughan, J., Daumé, H., III, Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).

Institut national de la statistique et des études économiques (INSEE). (2024, March 12). Estimations de population par sexe et âge au 1er janvier 2024: Comparaisons régionales et départementales. Retrieved February 27, 2025, from https://www.insee.fr/fr/statistiques/2012692.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131–164. https://doi.org/10.1037/rev0000093

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction, 7*, 1–29.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., … Hale, S. A. (2024). The PRISM alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models (arXiv:2404.16019). *arXiv*. https://doi.org/10.48550/arXiv.2404.16019

Lazar, S. (2024). Automatic authorities: Power and AI (arXiv:2404.05990). *arXiv*. Doi: 10.48550/arXiv.2404.05990.

Lazar, S., & Nelson, A. (2023). AI safety on whose terms? *Science, 381*(6654), 138. https://doi.org/10.1126/science.adi8982

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117–124).

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., … Clark, J. (2024). *The AI index 2024 annual report. AI Index Steering Committee, Institute for Human-Centered AI*. Stanford University. Retrieved 26 February 2025, from https://aiindex.stanford.edu/report/.

Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., … Hendrycks, D. (2025). Utility engineering: analyzing and controlling emergent value systems in AIs. *arXiv preprint*. arXiv:2502.08640.

McMillan, R., & Seetharaman, D. (2023). How a fervent belief Split Silicon Valley—And fueled the blowup at OpenAI. *WSJ*. https://www.wsj.com/tech/ai/openai-blowup-effective-altruism-disaster-f46a55e8.

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Nielsen, M. W., Alegria, S., Börjeson, L., Etzkowitz, H., Falk-Krzesinski, H. J., Joshi, A., … Schiebinger, L. (2017). Gender diversity leads to better science. *Proceedings of the National Academy of Sciences, 114*(8), 1740–1742. https://doi.org/10.1073/pnas.1700616114

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

Pal, S., Lazzaroni, R. M., & Mendoza, P. (2024). AI's missing link: The gender gap in the talent pool. *Interface*. Retrieved 26 February 2025, from https://www.interface-eu.org/publications/ai-gender-gap.

Pinker, S. (2011). *The better angels of our nature: The decline of violence in history and its causes*. Penguin UK.

Plaut, V. C., Garnett, F. G., Buffardi, L. E., & Sanchez-Burks, J. (2011). "What about me?" perceptions of exclusion and whites' reactions to multiculturalism. *Journal of Personality and Social Psychology, 101*(2), 337–353. https://doi.org/10.1037/a0022832

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

Purcell, Z. A., Dong, M., Nussberger, A.-M., Köbis, N., & Jakesch, M. (2024). People have different expectations for their own versus others' use of AI-mediated communication tools. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12727

Raji, I. D., & Buolamwini, J. (2019, January). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 429-435)*.

Rozado, D. (2024). The political preferences of LLMs (arXiv:2402.01789). *arXiv*. Doi: 10.48550/arXiv.2402.01789.

Sartori, L., & Theodorou, A. (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics and Information Technology, 24*(1), 4. https://doi.org/10.1007/s10676-022-09624-3

Scherrer, N., Shi, C., Feder, A., & Blei, D. M. (2023). Evaluating the moral beliefs encoded in LLMs (arXiv:2307.14324). *arXiv*. Doi: 10.48550/arXiv.2307.14324.

Schmitt, M. T., Branscombe, N. R., Postmes, T., & Garcia, A. (2014). The consequences of perceived discrimination for psychological well-being: A meta-analytic review. *Psychological Bulletin, 140*(4), 921–948. https://doi.org/10.1037/a0035754

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics, 9*(3), 331–352. https://doi.org/10.1007/s10892-005-3508-y

Takemoto, K. (2024). The moral machine experiment on large language models. *Royal Society Open Science, 11*(2), Article 231393. https://doi.org/10.1098/rsos.231393

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., … Gabriel, I. (2021). Ethical and social risks of harm from language models (arXiv:2112.04359). *arXiv*. Doi: 10.48550/arXiv.2112.04359.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., … Isaac, W. (2023). Sociotechnical safety evaluation of generative AI systems (arXiv:2310.11986). *arXiv*. Doi: 10.48550/arXiv.2310.11986.

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating systems: Gender, race and power in AI*. AI Now Institute. https://ainowinstitute.org/wp-content/uploads/2023/04/discriminatingsystems.pdf.