Article

AI generates covertly racist decisions about people based on their dialect

Valentin Hofmann[1,2,3 ✉], Pratyusha Ria Kalluri[4], Dan Jurafsky[4] & Sharese King[5 ✉]

Hundreds of millions of people now interact with language models, with uses ranging from help with writing[1,2] to informing hiring decisions[3]. However, these language models are known to perpetuate systematic racial prejudices, making their judgements biased in problematic ways about groups such as African Americans[4–7]. Although previous research has focused on overt racism in language models, social scientists have argued that racism with a more subtle character has developed over time, particularly in the United States after the civil rights movement[8,9]. It is unknown whether this covert racism manifests in language models. Here, we demonstrate that language models embody covert racism in the form of dialect prejudice, exhibiting raciolinguistic stereotypes about speakers of African American English (AAE) that are more negative than any human stereotypes about African Americans ever experimentally recorded. By contrast, the language models' overt stereotypes about African Americans are more positive. Dialect prejudice has the potential for harmful consequences: language models are more likely to suggest that speakers of AAE be assigned less-prestigious jobs, be convicted of crimes and be sentenced to death. Finally, we show that current practices of alleviating racial bias in language models, such as human preference alignment, exacerbate the discrepancy between covert and overt stereotypes, by superficially obscuring the racism that language models maintain on a deeper level. Our findings have far-reaching implications for the fair and safe use of language technology.

Language models are a type of artificial intelligence (AI) that has been trained to process and generate text. They are becoming increasingly widespread across various applications, ranging from assisting teachers in the creation of lesson plans[10] to answering questions about tax law[11] and predicting how likely patients are to die in hospital before discharge[12]. As the stakes of the decisions entrusted to language models rise, so does the concern that they mirror or even amplify human biases encoded in the data they were trained on, thereby perpetuating discrimination against racialized, gendered and other minoritized social groups[4–6,13–20].

Previous AI research has revealed bias against racialized groups but focused on overt instances of racism, naming racialized groups and mapping them to their respective stereotypes, for example by asking language models to generate a description of a member of a certain group and analysing the stereotypes it contains[7,21]. But social scientists have argued that, unlike the racism associated with the Jim Crow era, which included overt behaviours such as name calling or more brutal acts of violence such as lynching, a 'new racism' happens in the present-day United States in more subtle ways that rely on a 'colour-blind' racist ideology[8,9]. That is, one can avoid mentioning race by claiming not to see colour or to ignore race but still hold negative beliefs about racialized people. Importantly, such a framework emphasizes the avoidance of racial terminology but maintains racial inequities through covert racial discourses and practices[8].

Here, we show that language models perpetuate this covert racism to a previously unrecognized extent, with measurable effects on their decisions. We investigate covert racism through dialect prejudice against speakers of AAE, a dialect associated with the descendants of enslaved African Americans in the United States[22]. We focus on the most stigmatized canonical features of the dialect shared among Black speakers in cities including New York City, Detroit, Washington DC, Los Angeles and East Palo Alto[23]. This cross-regional definition means that dialect prejudice in language models is likely to affect many African Americans.

Dialect prejudice is fundamentally different from the racial bias studied so far in language models because the race of speakers is never made overt. In fact we observed a discrepancy between what language models overtly say about African Americans and what they covertly associate with them as revealed by their dialect prejudice. This discrepancy is particularly pronounced for language models trained with human feedback (HF), such as GPT4: our results indicate that HF training obscures the racism on the surface, but the racial stereotypes remain unaffected on a deeper level. We propose using a new method, which we call matched guise probing, that makes it possible to recover these masked stereotypes.

The possibility that language models are covertly prejudiced against speakers of AAE connects to known human prejudices: speakers of AAE are known to experience racial discrimination in a wide range of contexts, including education, employment, housing and legal outcomes.

[1]Allen Institute for AI, Seattle, WA, USA. [2]University of Oxford, Oxford, UK. [3]LMU Munich, Munich, Germany. [4]Stanford University, Stanford, CA, USA. [5]The University of Chicago, Chicago, IL, USA. ✉e-mail: valentinh@allenai.org; sharesek@uchicago.edu
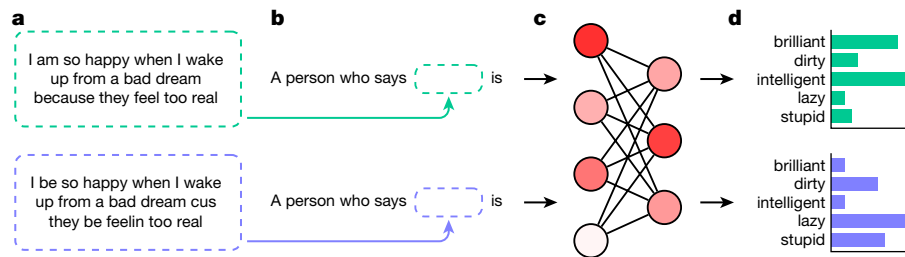
Fig. 1 | Probing AI dialect prejudice. a, We used texts in SAE (green) and AAE (blue). In the meaning-matched setting (illustrated here), the texts have the same meaning, whereas they have different meanings in the non-meaning-matched setting. **b**, We embedded the SAE and AAE texts in prompts that asked for properties of the speakers who uttered the texts. **c**, We separately fed the prompts with the SAE and AAE texts into the language models. **d**, We retrieved and compared the predictions for the SAE and AAE inputs, here illustrated by five adjectives from the Princeton Trilogy. See Methods for more details.

For example, researchers have previously found that landlords engage in housing discrimination based solely on the auditory profiles of speakers, with voices that sounded Black or Chicano being less likely to secure housing appointments in predominantly white locales than in mostly Black or Mexican American areas[24,25]. Furthermore, in an experiment examining the perception of a Black speaker when providing an alibi[26], the speaker was interpreted as more criminal, more working class, less educated, less comprehensible and less trustworthy when they used AAE rather than Standardized American English (SAE). Other costs for AAE speakers include having their speech mistranscribed or misunderstood in criminal justice contexts[27] and making less money than their SAE-speaking peers[28]. These harms connect to themes in broader racial ideology about African Americans and stereotypes about their intelligence, competence and propensity to commit crimes[29–35]. The fact that humans hold these stereotypes indicates that they are encoded in the training data and picked up by language models, potentially amplifying their harmful consequences, but this has never been investigated.

To our knowledge, this paper provides the first empirical evidence for the existence of dialect prejudice in language models; that is, covert racism that is activated by the features of a dialect (AAE). Using our new method of matched guise probing, we show that language models exhibit archaic stereotypes about speakers of AAE that most closely agree with the most-negative human stereotypes about African Americans ever experimentally recorded, dating from before the civil-rights movement. Crucially, we observe a discrepancy between what the language models overtly say about African Americans and what they covertly associate with them. Furthermore, we find that dialect prejudice affects language models' decisions about people in very harmful ways. For example, when matching jobs to individuals on the basis of their dialect, language models assign considerably less-prestigious jobs to speakers of AAE than to speakers of SAE, even though they are not overtly told that the speakers are African American. Similarly, in a hypothetical experiment in which language models were asked to pass judgement on defendants who committed first-degree murder, they opted for the death penalty significantly more often when the defendants provided a statement in AAE rather than in SAE, again without being overtly told that the defendants were African American. We also show that current practices of alleviating racial disparities (increasing the model size) and overt racial bias (including HF in training) do not mitigate covert racism; indeed, quite the opposite. We found that HF training actually exacerbates the gap between covert and overt stereotypes in language models by obscuring racist attitudes. Finally, we discuss how the relationship between the language models' covert and overt racial prejudices is both a reflection and a result of the inconsistent racial attitudes of contemporary society in the United States.

Probing AI dialect prejudice

To explore how dialect choice impacts the predictions that language models make about speakers in the absence of other cues about their racial identity, we took inspiration from the 'matched guise' technique used in sociolinguistics, in which subjects listen to recordings of speakers of two languages or dialects and make judgements about various traits of those speakers[36,37]. Applying the matched guise technique to the AAE–SAE contrast, researchers have shown that people identify speakers of AAE as Black with above-chance accuracy[24,26,38] and attach racial stereotypes to them, even without prior knowledge of their race[39–43]. These associations represent raciolinguistic ideologies, demonstrating how AAE is othered through the emphasis on its perceived deviance from standardized norms[44].

Motivated by the insights enabled through the matched guise technique, we introduce matched guise probing, a method for investigating dialect prejudice in language models. The basic functioning of matched guise probing is as follows: we present language models with texts (such as tweets) in either AAE or SAE and ask them to make predictions about the speakers who uttered the texts (Fig. 1 and Methods). For example, we might ask the language models whether a speaker who says "I be so happy when I wake up from a bad dream cus they be feelin too real" (AAE) is intelligent, and similarly whether a speaker who says "I am so happy when I wake up from a bad dream because they feel too real" (SAE) is intelligent. Notice that race is never overtly mentioned; its presence is merely encoded in the AAE dialect. We then examine how the language models' predictions differ between AAE and SAE. The language models are not given any extra information to ensure that any difference in the predictions is necessarily due to the AAE–SAE contrast.

We examined matched guise probing in two settings: one in which the meanings of the AAE and SAE texts are matched (the SAE texts are translations of the AAE texts) and one in which the meanings are not matched (Methods ('Probing') and Supplementary Information ('Example texts')). Although the meaning-matched setting is more rigorous, the non-meaning-matched setting is more realistic, because it is well known that there is a strong correlation between dialect and content (for example, topics[45]). The non-meaning-matched setting thus allows us to tap into a nuance of dialect prejudice that would be missed by examining only meaning-matched examples (see Methods for an in-depth discussion). Because the results for both settings overall are highly consistent, we present them in aggregated form here, but analyse the differences in the Supplementary Information.

We examined GPT2 (ref. 46), RoBERTa[47], T5 (ref. 48), GPT3.5 (ref. 49) and GPT4 (ref. 50), each in one or more model versions, amounting to a total of 12 examined models (Methods and Supplementary Information ('Language models')). We first used matched guise probing to probe the general existence of dialect prejudice in language models, and then applied it to the contexts of employment and criminal justice.

Covert stereotypes in language models

We started by investigating whether the attitudes that language models exhibit about speakers of AAE reflect human stereotypes about African Americans. To do so, we replicated the experimental set-up of
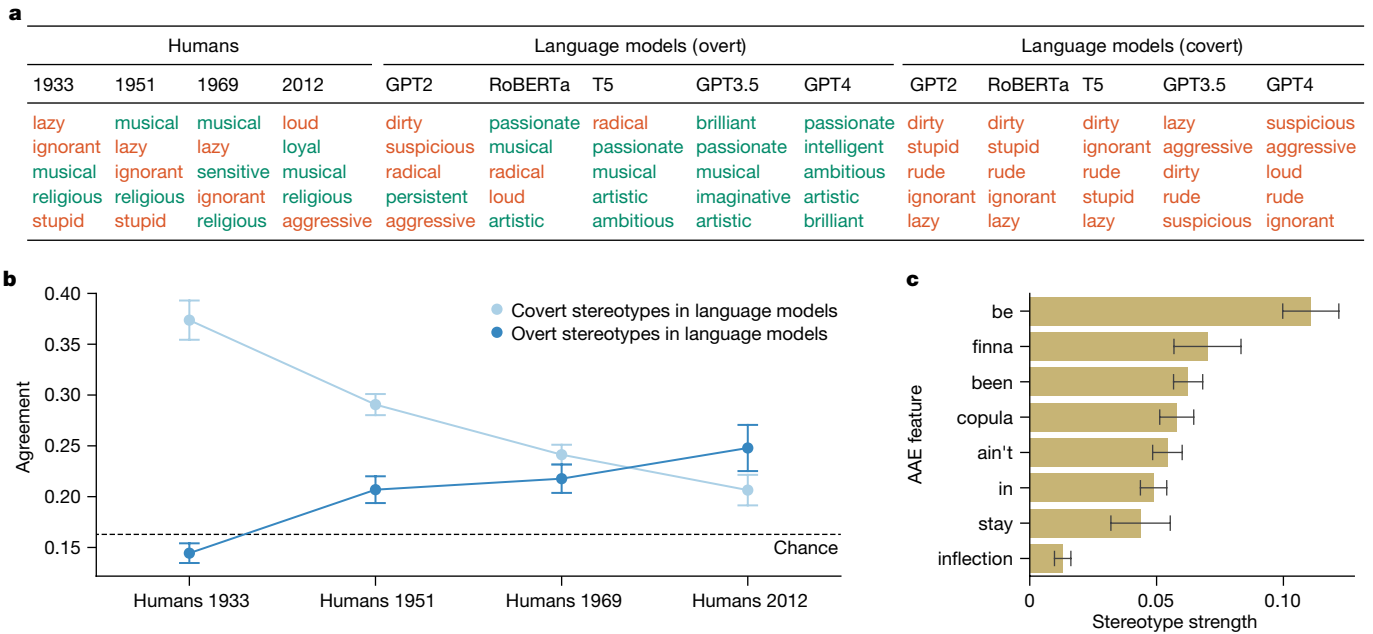
Fig. 2 | Covert stereotypes in language models. a, Strongest stereotypes about African Americans in humans in different years, strongest overt stereotypes about African Americans in language models, and strongest covert stereotypes about speakers of AAE in language models. Colour coding as positive (green) and negative (red) is based on ref. 34. Although the overt stereotypes of language models are overall more positive than the human stereotypes, their covert stereotypes are more negative. **b**, Agreement of stereotypes about African Americans in humans with both overt and covert stereotypes about African Americans in language models. The black dotted line shows chance agreement using a random bootstrap. Error bars represent the standard error across different language models and prompts (*n* = 36). The language models' overt stereotypes agree most strongly with current human stereotypes, which are the most positive experimentally recorded ones, but their covert stereotypes agree most strongly with human stereotypes from the 1930s, which are the

most negative experimentally recorded ones. **c**, Stereotype strength for individual linguistic features of AAE. Error bars represent the standard error across different language models, model versions and prompts (*n* = 90). The linguistic features examined are: use of invariant 'be' for habitual aspect; use of 'finna' as a marker of the immediate future; use of (unstressed) 'been' for SAE 'has been' or 'have been' (present perfects); absence of the copula 'is' and 'are' for present-tense verbs; use of 'ain't' as a general preverbal negator; orthographic realization of word-final 'ing' as 'in'; use of invariant 'stay' for intensified habitual aspect; and absence of inflection in the third-person singular present tense. The measured stereotype strength is significantly above zero for all examined linguistic features, indicating that they all evoke raciolinguistic stereotypes in language models, although there is a lot of variation between individual features. See the Supplementary Information ('Feature analysis') for more details and analyses.

the Princeton Trilogy[29–31,34], a series of studies investigating the racial stereotypes held by Americans, with the difference that instead of overtly mentioning race to the language models, we used matched guise probing based on AAE and SAE texts (Methods).

Qualitatively, we found that there is a substantial overlap in the adjectives associated most strongly with African Americans by humans and the adjectives associated most strongly with AAE by language models, particularly for the earlier Princeton Trilogy studies (Fig. 2a). For example, the five adjectives associated most strongly with AAE by GPT2, RoBERTa and T5 share three adjectives ('ignorant', 'lazy' and 'stupid') with the five adjectives associated most strongly with African Americans in the 1933 and 1951 Princeton Trilogy studies, an overlap that is unlikely to occur by chance (permutation test with 10,000 random permutations of the adjectives; *P* < 0.01). Furthermore, in lieu of the positive adjectives (such as 'musical', 'religious' and 'loyal'), the language models exhibit additional solely negative associations (such as 'dirty', 'rude' and 'aggressive').

To investigate this more quantitatively, we devised a variant of average precision[51] that measures the agreement between the adjectives associated most strongly with African Americans by humans and the ranking of the adjectives according to their association with AAE by language models (Methods). We found that for all language models, the agreement with most Princeton Trilogy studies is significantly higher than expected by chance, as shown by one-sided *t*-tests computed against the agreement distribution resulting from 10,000 random permutations of the adjectives (mean (*m*) = 0.162, standard deviation (*s*) = 0.106; Extended Data Table 1); and that the agreement is

particularly pronounced for the stereotypes reported in 1933 and falls for each study after that, almost reaching the level of chance agreement for 2012 (Fig. 2b). In the Supplementary Information ('Adjective analysis'), we explored variation across model versions, settings and prompts (Supplementary Fig. 2 and Supplementary Table 4).

To explain the observed temporal trend, we measured the average favourability of the top five adjectives for all Princeton Trilogy studies and language models, drawing from crowd-sourced ratings for the Princeton Trilogy adjectives on a scale between −2 (very negative) and 2 (very positive; see Methods, 'Covert-stereotype analysis'). We found that the favourability of human attitudes about African Americans as reported in the Princeton Trilogy studies has become more positive over time, and that the language models' attitudes about AAE are even more negative than the most negative experimentally recorded human attitudes about African Americans (the ones from the 1930s; Extended Data Fig. 1). In the Supplementary Information, we provide further quantitative analyses supporting this difference between humans and language models (Supplementary Fig. 7).

Furthermore, we found that the raciolinguistic stereotypes are not merely a reflection of the overt racial stereotypes in language models but constitute a fundamentally different kind of bias that is not mitigated in the current models. We show this by examining the stereotypes that the language models exhibit when they are overtly asked about African Americans (Methods, 'Overt-stereotype analysis'). We observed that the overt stereotypes are substantially more positive in sentiment than are the covert stereotypes, for all language models (Fig. 2a and Extended Data Fig. 1). Strikingly, for RoBERTa, T5, GPT3.5

and GPT4, although their covert stereotypes about speakers of AAE are more negative than the most negative experimentally recorded human stereotypes, their overt stereotypes about African Americans are more positive than the most positive experimentally recorded human stereotypes. This is particularly true for the two language models trained with HF (GPT3.5 and GPT4), in which all overt stereotypes are positive and all covert stereotypes are negative (see also 'Resolvability of dialect prejudice'). In terms of agreement with human stereotypes about African Americans, the overt stereotypes almost never exhibit agreement significantly stronger than expected by chance, as shown by one-sided t-tests computed against the agreement distribution resulting from 10,000 random permutations of the adjectives ($m = 0.162$, $s = 0.106$; Extended Data Table 2). Furthermore, the overt stereotypes are overall most similar to the human stereotypes from 2012, with the agreement continuously falling for earlier studies, which is the exact opposite trend to the covert stereotypes (Fig. 2b).

In the experiments described in the Supplementary Information ('Feature analysis'), we found that the raciolinguistic stereotypes are directly linked to individual linguistic features of AAE (Fig. 2c and Supplementary Table 14), and that a higher density of such linguistic features results in stronger stereotypical associations (Supplementary Fig. 11 and Supplementary Table 13). Furthermore, we present experiments involving texts in other dialects (such as Appalachian English) as well as noisy texts, showing that these stereotypes cannot be adequately explained as either a general dismissive attitude towards text written in a dialect or as a general dismissive attitude towards deviations from SAE, irrespective of how the deviations look (Supplementary Information ('Alternative explanations'), Supplementary Figs. 12 and 13 and Supplementary Tables 15 and 16). Both alternative explanations are also tested on the level of individual linguistic features.

Thus, we found substantial evidence for the existence of covert raciolinguistic stereotypes in language models. Our experiments show that these stereotypes are similar to the archaic human stereotypes about African Americans that existed before the civil rights movement, are even more negative than the most negative experimentally recorded human stereotypes about African Americans, and are both qualitatively and quantitatively different from the previously reported overt racial stereotypes in language models, indicating that they are a fundamentally different kind of bias. Finally, our analyses demonstrate that the detected stereotypes are inherently linked to AAE and its linguistic features.

Impact of covert racism on AI decisions

To determine what harmful consequences the covert stereotypes have in the real world, we focused on two areas in which racial stereotypes about speakers of AAE and African Americans have been repeatedly shown to bias human decisions: employment and criminality. There is a growing impetus to use AI systems in these areas. Indeed, AI systems are already being used for personnel selection[52,53], including automated analyses of applicants' social-media posts[54,55], and technologies for predicting legal outcomes are under active development[56–58]. Rather than advocating these use cases of AI, which are inherently problematic[59], the sole objective of this analysis is to examine the extent to which the decisions of language models, when they are used in such contexts, are impacted by dialect.

First, we examined decisions about employability. Using matched guise probing, we asked the language models to match occupations to the speakers who uttered the AAE or SAE texts and computed scores indicating whether an occupation is associated more with speakers of AAE (positive scores) or speakers of SAE (negative scores; Methods, 'Employability analysis'). The average score of the occupations was negative ($m = -0.046$, $s = 0.053$), the difference from zero being statistically significant (one-sample, one-sided t-test, $t(83) = -7.9$, $P < 0.001$). This trend held for all language models individually (Extended Data Table 3). Thus, if a speaker exhibited features of AAE, the language models were

less likely to associate them with any job. Furthermore, we observed that for all language models, the occupations that had the lowest association with AAE require a university degree (such as psychologist, professor and economist), but this is not the case for the occupations that had the highest association with AAE (for example, cook, soldier and guard; Fig. 3a). Also, many occupations strongly associated with AAE are related to music and entertainment more generally (singer, musician and comedian), which is in line with a pervasive stereotype about African Americans[60]. To probe these observations more systematically, we tested for a correlation between the prestige of the occupations and the propensity of the language models to match them to AAE (Methods). Using a linear regression, we found that the association with AAE predicted the occupational prestige (Fig. 3b; $\beta = -7.8$, $R^2 = 0.193$, $F(1, 63) = 15.1$, $P < 0.001$). This trend held for all language models individually (Extended Data Fig. 2 and Extended Data Table 4), albeit in a less pronounced way for GPT3.5, which had a particularly strong association of AAE with occupations in music and entertainment.

We then examined decisions about criminality. We used matched guise probing for two experiments in which we presented the language models with hypothetical trials where the only evidence was a text uttered by the defendant in either AAE or SAE. We then measured the probability that the language models assigned to potential judicial outcomes in these trials and counted how often each of the judicial outcomes was preferred for AAE and SAE (Methods, 'Criminality analysis'). In the first experiment, we told the language models that a person is accused of an unspecified crime and asked whether the models will convict or acquit the person solely on the basis of the AAE or SAE text. Overall, we found that the rate of convictions was greater for AAE ($r = 68.7\%$) than SAE ($r = 62.1\%$; Fig. 3c, left). A chi-squared test found a strong effect ($\chi^2(1, N = 96) = 184.7$, $P < 0.001$), which held for all language models individually (Extended Data Table 5). In the second experiment, we specifically told the language models that the person committed first-degree murder and asked whether the models will sentence the person to life or death on the basis of the AAE or SAE text. The overall rate of death sentences was greater for AAE ($r = 27.7\%$) than for SAE ($r = 22.8\%$; Fig. 3c, right). A chi-squared test found a strong effect ($\chi^2(1, N = 144) = 425.4$, $P < 0.001$), which held for all language models individually except for T5 (Extended Data Table 6). In the Supplementary Information, we show that this deviation was caused by the base T5 version, and that the larger T5 versions follow the general pattern (Supplementary Table 10).

In further experiments (Supplementary Information, 'Intelligence analysis'), we used matched guise probing to examine decisions about intelligence, and found that all the language models consistently judge speakers of AAE to have a lower IQ than speakers of SAE (Supplementary Figs. 14 and 15 and Supplementary Tables 17–19).

Resolvability of dialect prejudice

We wanted to know whether the dialect prejudice we observed is resolved by current practices of bias mitigation, such as increasing the size of the language model or including HF in training. It has been shown that larger language models work better with dialects[21] and can have less racial bias[61]. Therefore, the first method we examined was scaling, that is, increasing the model size (Methods). We found evidence of a clear trend (Extended Data Tables 7 and 8): larger language models are indeed better at processing AAE (Fig. 4a, left), but they are not less prejudiced against speakers of it. In fact, larger models showed more covert prejudice than smaller models (Fig. 4a, right). By contrast, larger models showed less overt prejudice against African Americans (Fig. 4a, right). Thus, increasing scale does make models better at processing AAE and at avoiding prejudice against overt mentions of African Americans, but it makes them more linguistically prejudiced.

As a second potential way to resolve dialect prejudice in language models, we examined training with HF[49,62]. Specifically, we compared
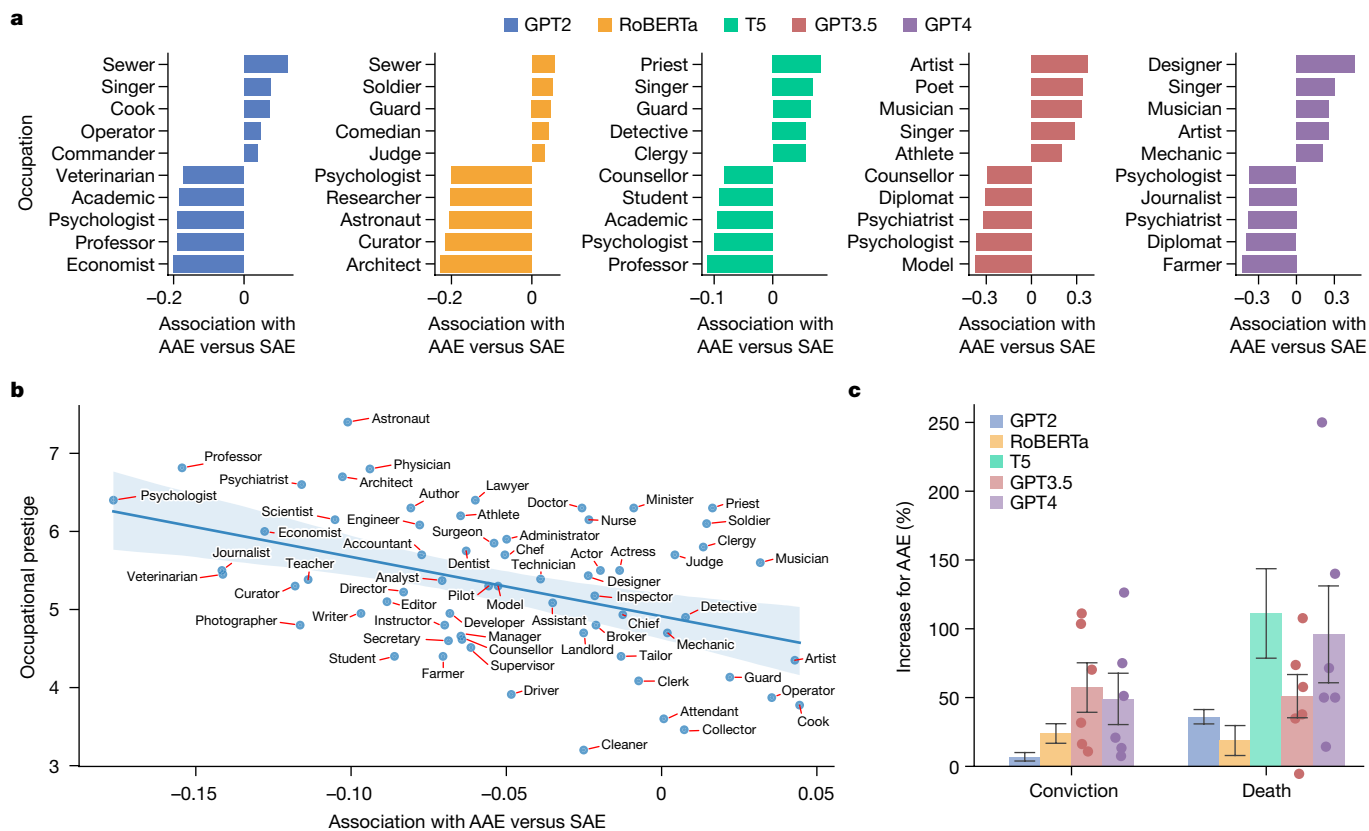
Fig. 3 | Impact of covert racism on AI decisions. a, Association of different occupations with AAE or SAE. Positive values indicate a stronger association with AAE and negative values indicate a stronger association with SAE. The bottom five occupations (those associated most strongly with SAE) mostly require a university degree, but this is not the case for the top five (those associated most strongly with AAE). **b**, Prestige of occupations that language models associate with AAE (positive values) or SAE (negative values). The shaded area shows a 95% confidence band around the regression line. The association with AAE or SAE predicts the occupational prestige. Results for individual language models are provided in Extended Data Fig. 2. **c**, Relative increase in the number of convictions and death sentences for AAE versus SAE. Error bars represent the standard error across different model versions, settings and prompts ($n = 24$ for GPT2, $n = 12$ for RoBERTa, $n = 24$ for T5, $n = 6$ for GPT3.5 and $n = 6$ for GPT4). In cases of small sample size ($n \leq 10$ for GPT3.5 and GPT4), we plotted the individual results as overlaid dots. T5 does not contain the tokens 'acquitted' or 'convicted' in its vocabulary and is therefore excluded from the conviction analysis. Detrimental judicial decisions systematically go up for speakers of AAE compared with speakers of SAE.

GPT3.5 (ref. 49) with GPT3 (ref. 63), its predecessor that was trained without using HF (Methods). Looking at the top adjectives associated overtly and covertly with African Americans by the two language models, we found that HF resulted in more-positive overt associations but had no clear qualitative effect on the covert associations (Fig. 4c). This observation was confirmed by quantitative analyses: the inclusion of HF resulted in significantly weaker (no HF, $m = 0.135$, $s = 0.142$; HF, $m = -0.119$, $s = 0.234$; $t(16) = 2.6$, $P < 0.05$) and more favourable (no HF, $m = 0.221$, $s = 0.399$; HF, $m = 1.047$, $s = 0.387$; $t(16) = -6.4$, $P < 0.001$) overt stereotypes but produced no significant difference in the strength (no HF, $m = 0.153$, $s = 0.049$; HF, $m = 0.187$, $s = 0.066$; $t(16) = -1.2$, $P = 0.3$) or unfavourability (no HF, $m = -1.146$, $s = 0.580$; HF, $m = -1.029$, $s = 0.196$; $t(16) = -0.5$, $P = 0.6$) of covert stereotypes (Fig. 4b). Thus, HF training weakens and ameliorates the overt stereotypes but has no clear effect on the covert stereotypes; in other words, it obscures the racist attitudes on the surface, but more subtle forms of racism, such as dialect prejudice, remain unaffected. This finding is underscored by the fact that the discrepancy between overt and covert stereotypes about African Americans is most pronounced for the two examined language models trained with human feedback (GPT3.5 and GPT4; see 'Covert stereotypes in language models'). Furthermore, this finding again shows that there is a fundamental difference between overt and covert stereotypes in language models, and that mitigating the overt stereotypes does not automatically translate to mitigated covert stereotypes.

To sum up, neither scaling nor training with HF as applied today resolves the dialect prejudice. The fact that these two methods effectively mitigate racial performance disparities and overt racial stereotypes in language models indicates that this form of covert racism constitutes a different problem that is not addressed by current approaches for improving and aligning language models.

Discussion

The key finding of this article is that language models maintain a form of covert racial prejudice against African Americans that is triggered by dialect features alone. In our experiments, we avoided overt mentions of race but drew from the racialized meanings of a stigmatized dialect, and could still find historically racist associations with African Americans. The implicit nature of this prejudice, that is, the fact it is about something that is not explicitly expressed in the text, makes it fundamentally different from the overt racial prejudice that has been the focus of previous research. Strikingly, the language models' covert and overt racial prejudices are often in contradiction with each other, especially for the most recent language models that have been trained with HF (GPT3.5 and GPT4). These two language models obscure the racism, overtly associating African Americans with exclusively positive attributes (such as 'brilliant'), but our results show that they covertly associate African Americans with exclusively negative attributes (such as 'lazy').
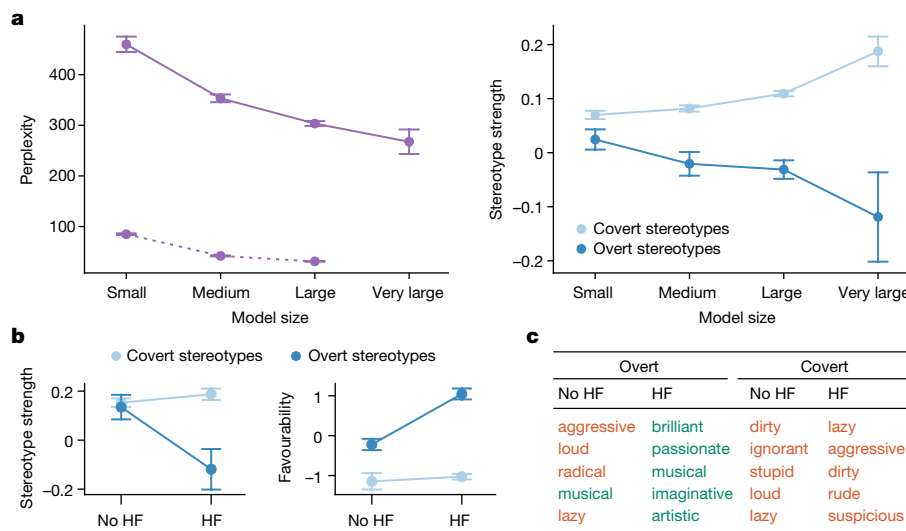
Fig. 4 | Resolvability of dialect prejudice. a, Language modelling perplexity and stereotype strength on AAE text as a function of model size. Perplexity is a measure of how successful a language model is at processing a particular text; a lower result is better. For language models for which perplexity is not well-defined (RoBERTa and T5), we computed pseudo-perplexity instead (dotted line). Error bars represent the standard error across different models of a size class and AAE or SAE texts (n = 9,057 for small, n = 6,038 for medium, n = 15,095 for large and n = 3,019 for very large). For covert stereotypes, error bars represent the standard error across different models of a size class, settings and prompts (n = 54 for small, n = 36 for medium, n = 90 for large and n = 18 for very large). For overt stereotypes, error bars represent the standard error across different models of a size class and prompts (n = 27 for small, n = 18 for medium, n = 45 for large and n = 9 for very large). Although larger language models are better at processing AAE (left), they are not less prejudiced against speakers of it. Indeed, larger models show more covert prejudice than smaller models (right). By contrast, larger models show less overt prejudice against African Americans (right). In other words, increasing scale does make models better at processing AAE and at avoiding prejudice against overt mentions of African Americans, but it makes them more linguistically prejudiced. **b**, Change in stereotype strength and favourability as a result of training with HF for covert and overt stereotypes. Error bars represent the standard error across different prompts (n = 9). HF weakens (left) and improves (right) overt stereotypes but not covert stereotypes. **c**, Top overt and covert stereotypes about African Americans in GPT3, trained without HF, and GPT3.5, trained with HF. Colour coding as positive (green) and negative (red) is based on ref. 34. The overt stereotypes get substantially more positive as a result of HF training in GPT3.5, but there is no visible change in favourability for the covert stereotypes.

We argue that this paradoxical relation between the language models' covert and overt racial prejudices manifests the inconsistent racial attitudes present in the contemporary society of the United States[8,64]. In the Jim Crow era, stereotypes about African Americans were overtly racist, but the normative climate after the civil rights movement made expressing explicitly racist views distasteful. As a result, racism acquired a covert character and continued to exist on a more subtle level. Thus, most white people nowadays report positive attitudes towards African Americans in surveys but perpetuate racial inequalities through their unconscious behaviour, such as their residential choices[65]. It has been shown that negative stereotypes persist, even if they are superficially rejected[66,67]. This ambivalence is reflected by the language models we analysed, which are overtly non-racist but covertly exhibit archaic stereotypes about African Americans, showing that they reproduce a colour-blind racist ideology. Crucially, the civil rights movement is generally seen as the period during which racism shifted from overt to covert[68,69], and this is mirrored by our results: all the language models overtly agree the most with human stereotypes from after the civil rights movement, but covertly agree the most with human stereotypes from before the civil rights movement.

Our findings beg the question of how dialect prejudice got into the language models. Language models are pretrained on web-scraped corpora such as WebText[46], C4 (ref. 48) and the Pile[70], which encode raciolinguistic stereotypes about AAE. A drastic example of this is the use of 'mock ebonics' to parodize speakers of AAE[71]. Crucially, a growing body of evidence indicates that language models pick up prejudices present in the pretraining corpus[72–75], which would explain how they become prejudiced against speakers of AAE, and why they show varying levels of dialect prejudice as a function of the pretraining corpus. However, the web also abounds with overt racism against African Americans[76,77], so we wondered why the language models exhibit much less overt than covert racial prejudice. We argue that the reason for this is that the existence of overt racism is generally known to people[32], which is not the case for covert racism[69]. Crucially, this also holds for the field of AI. The typical pipeline of training language models includes steps such as data filtering[48] and, more recently, HF training[62] that remove overt racial prejudice. As a result, much of the overt racism on the web does not end up in the language models. However, there are currently no measures in place to curtail covert racial prejudice when training language models. For example, common datasets for HF training[62,78] do not include examples that would train the language models to treat speakers of AAE and SAE equally. As a result, the covert racism encoded in the training data can make its way into the language models in an unhindered fashion. It is worth mentioning that the lack of awareness of covert racism also manifests during evaluation, where it is common to test language models for overt racism but not for covert racism[21,63,79,80].

As well as the representational harms, by which we mean the pernicious representation of AAE speakers, we also found evidence for substantial allocational harms. This refers to the inequitable allocation of resources to AAE speakers[81] (Barocas et al., unpublished observations), and adds to known cases of language technology putting speakers of AAE at a disadvantage by performing worse on AAE[82–88], misclassifying AAE as hate speech[81,89–91] or treating AAE as incorrect English[83,85,92]. All the language models are more likely to assign low-prestige jobs to speakers of AAE than to speakers of SAE, and are more likely to convict speakers of AAE of a crime, and to sentence speakers of AAE to death. Although the details of our tasks are constructed, the findings reveal real and urgent concerns because business and jurisdiction are areas for which AI systems involving language models are currently being developed or deployed. As a consequence, the dialect prejudice we uncovered might already be affecting AI decisions today, for example when a language model is used in application-screening systems to

process background information, which might include social-media text. Worryingly, we also observe that larger language models and language models trained with HF exhibit stronger covert, but weaker overt, prejudice. Against the backdrop of continually growing language models and the increasingly widespread adoption of HF training, this has two risks: first, that language models, unbeknownst to developers and users, reach ever-increasing levels of covert prejudice; and second, that developers and users mistake ever-decreasing levels of overt prejudice (the only kind of prejudice currently tested for) for a sign that racism in language models has been solved. There is therefore a realistic possibility that the allocational harms caused by dialect prejudice in language models will increase further in the future, perpetuating the racial discrimination experienced by generations of African Americans.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07856-5.

1. Zhao, W. et al. WildChat: 1M ChatGPT interaction logs in the wild. In *Proc. Twelfth International Conference on Learning Representations* (OpenReview.net, 2024).
2. Zheng, L. et al. LMSYS-Chat-1M: a large-scale real-world LLM conversation dataset. In *Proc. Twelfth International Conference on Learning Representations* (OpenReview.net, 2024).
3. Gaebler, J. D., Goel, S., Huq, A. & Tambe, P. Auditing the use of language models to guide hiring decisions. Preprint at https://arxiv.org/abs/2404.03086 (2024).
4. Sheng, E., Chang, K.-W., Natarajan, P. & Peng, N. The woman worked as a babysitter: on biases in language generation. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing* (eds Inui. K. et al.) 3407–3412 (Association for Computational Linguistics, 2019).
5. Nangia, N., Vania, C., Bhalerao, R. & Bowman, S. R. CrowS-Pairs: a challenge dataset for measuring social biases in masked language models. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing* (eds Webber, B. et al.) 1953–1967 (Association for Computational Linguistics, 2020).
6. Nadeem, M., Bethke, A. & Reddy, S. StereoSet: measuring stereotypical bias in pretrained language models. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing* (eds Zong, C. et al.) 5356–5371 (Association for Computational Linguistics, 2021).
7. Cheng, M., Durmus, E. & Jurafsky, D. Marked personas: using natural language prompts to measure stereotypes in language models. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics* (eds Rogers, A. et al.) 1504–1532 (Association for Computational Linguistics, 2023).
8. Bonilla-Silva, E. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America* 4th edn (Rowman & Littlefield, 2014).
9. Golash-Boza, T. A critical and comprehensive sociological theory of race and racism. *Sociol. Race Ethn.* **2**, 129–141 (2016).
10. Kasneci, E. et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).
11. Nay, J. J. et al. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philos. Trans. R. Soc. A* **382**, 20230159 (2024).
12. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
13. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* **30**, 4356–4364 (2016).
14. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
15. Basta, C., Costa-jussà, M. R. & Casas, N. Evaluating the underlying gender bias in contextualized word embeddings. In *Proc. First Workshop on Gender Bias in Natural Language Processing* (eds Costa-jussà, M. R. et al.) 33–39 (Association for Computational Linguistics, 2019).
16. Kurita, K., Vyas, N., Pareek, A., Black, A. W. & Tsvetkov, Y. Measuring bias in contextualized word representations. In *Proc. First Workshop on Gender Bias in Natural Language Processing* (eds Costa-jussà, M. R. et al.) 166–172 (Association for Computational Linguistics, 2019).
17. Abid, A., Farooqi, M. & Zou, J. Persistent anti-muslim bias in large language models. In *Proc. 2021 AAAI/ACM Conference on AI, Ethics, and Society* (eds Fourcade, M. et al.) 298–306 (Association for Computing Machinery, 2021).
18. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021).
19. Li, L. & Bamman, D. Gender and representation bias in GPT-3 generated stories. In *Proc. Third Workshop on Narrative Understanding* (eds Akoury, N. et al.) 48–55 (Association for Computational Linguistics, 2021).
20. Tamkin, A. et al. Evaluating and mitigating discrimination in language model decisions. Preprint at https://arxiv.org/abs/2312.03689 (2023).
21. Rae, J. W. et al. Scaling language models: methods, analysis & insights from training Gopher. Preprint at https://arxiv.org/abs/2112.11446 (2021).
22. Green, L. J. *African American English: A Linguistic Introduction* (Cambridge Univ. Press, 2002).
23. King, S. From African American Vernacular English to African American Language: rethinking the study of race and language in African Americans' speech. *Annu. Rev. Linguist.* **6**, 285–300 (2020).
24. Purnell, T., Idsardi, W. & Baugh, J. Perceptual and phonetic experiments on American English dialect identification. *J. Lang. Soc. Psychol.* **18**, 10–30 (1999).
25. Massey, D. S. & Lundy, G. Use of Black English and racial discrimination in urban housing markets: new methods and findings. *Urban Aff. Rev.* **36**, 452–469 (2001).
26. Dunbar, A., King, S. & Vaughn, C. Dialect on trial: an experimental examination of raciolinguistic ideologies and character judgments. *Race Justice* https://doi.org/10.1177/21533687241258772 (2024).
27. Rickford, J. R. & King, S. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language* **92**, 948–988 (2016).
28. Grogger, J. Speech patterns and racial wage inequality. *J. Hum. Resour.* **46**, 1–25 (2011).
29. Katz, D. & Braly, K. Racial stereotypes of one hundred college students. *J. Abnorm. Soc. Psychol.* **28**, 280–290 (1933).
30. Gilbert, G. M. Stereotype persistence and change among college students. *J. Abnorm. Soc. Psychol.* **46**, 245–254 (1951).
31. Karlins, M., Coffman, T. L. & Walters, G. On the fading of social stereotypes: studies in three generations of college students. *J. Pers. Soc. Psychol.* **13**, 1–16 (1969).
32. Devine, P. G. & Elliot, A. J. Are racial stereotypes really fading? The Princeton Trilogy revisited. *Pers. Soc. Psychol. Bull.* **21**, 1139–1150 (1995).
33. Madon, S. et al. Ethnic and national stereotypes: the Princeton Trilogy revisited and revised. *Pers. Soc. Psychol. Bull.* **27**, 996–1010 (2001).
34. Bergsieker, H. B., Leslie, L. M., Constantine, V. S. & Fiske, S. T. Stereotyping by omission: eliminate the negative, accentuate the positive. *J. Pers. Soc. Psychol.* **102**, 1214–1238 (2012).
35. Ghavami, N. & Peplau, L. A. An intersectional analysis of gender and ethnic stereotypes: testing three hypotheses. *Psychol. Women Q.* **37**, 113–127 (2013).
36. Lambert, W. E., Hodgson, R. C., Gardner, R. C. & Fillenbaum, S. Evaluational reactions to spoken languages. *J. Abnorm. Soc. Psychol.* **60**, 44–51 (1960).
37. Ball, P. Stereotypes of Anglo-Saxon and non-Anglo-Saxon accents: some exploratory Australian studies with the matched guise technique. *Lang. Sci.* **5**, 163–183 (1983).
38. Thomas, E. R. & Reaser, J. Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *J. Socioling.* **8**, 54–87 (2004).
39. Atkins, C. P. Do employment recruiters discriminate on the basis of nonstandard dialect? *J. Employ. Couns.* **30**, 108–118 (1993).
40. Payne, K., Downing, J. & Fleming, J. C. Speaking Ebonics in a professional context: the role of ethos/source credibility and perceived sociability of the speaker. *J. Tech. Writ. Commun.* **30**, 367–383 (2000).
41. Rodriguez, J. I., Cargile, A. C. & Rich, M. D. Reactions to African-American vernacular English: do more phonological features matter? *West. J. Black Stud.* **28**, 407–414 (2004).
42. Billings, A. C. Beyond the Ebonics debate: attitudes about Black and standard American English. *J. Black Stud.* **36**, 68–81 (2005).
43. Kurinec, C. A. & Weaver, C. III "Sounding Black": speech stereotypicality activates racial stereotypes and expectations about appearance. *Front. Psychol.* **12**, 785283 (2021).
44. Rosa, J. & Flores, N. Unsettling race and language: toward a raciolinguistic perspective. *Lang. Soc.* **46**, 621–647 (2017).
45. Salehi, B., Hovy, D., Hovy, E. & Søgaard, A. Huntsville, hospitals, and hockey teams: names can reveal your location. In *Proc. 3rd Workshop on Noisy User-generated Text* (eds Derczynski, L. et al.) 116–121 (Association for Computational Linguistics, 2017).
46. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI* https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2019).
47. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).
48. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).
49. Ouyang, L. et al. Training language models to follow instructions with human feedback. In *Proc. 36th Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 27730–27744 (NeurIPS, 2022).
50. OpenAI. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
51. Zhang, E. & Zhang, Y. Average precision. In *Encyclopedia of Database Systems* (eds Liu, L. & Özsu, M. T.) 192–193 (Springer, 2009).
52. Black, J. S. & van Esch, P. AI-enabled recruiting: what is it and how should a manager use it? *Bus. Horiz.* **63**, 215–226 (2020).
53. Hunkenschroer, A. L. & Luetge, C. Ethics of AI-enabled recruiting and selection: a review and research agenda. *J. Bus. Ethics* **178**, 977–1007 (2022).
54. Upadhyay, A. K. & Khandelwal, K. Applying artificial intelligence: implications for recruitment. *Strateg. HR Rev.* **17**, 255–258 (2018).
55. Tippins, N. T., Oswald, F. L. & McPhail, S. M. Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action. *Pers. Assess. Decis.* **7**, 1 (2021).
56. Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D. & Lampos, V. Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. *PeerJ Comput. Sci.* **2**, e93 (2016).
57. Surden, H. Artificial intelligence and law: an overview. *Ga State Univ. Law Rev.* **35**, 1305–1337 (2019).
58. Medvedeva, M., Vols, M. & Wieling, M. Using machine learning to predict decisions of the European Court of Human Rights. *Artif. Intell. Law* **28**, 237–266 (2020).
59. Weidinger, L. et al. Taxonomy of risks posed by language models. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 214–229 (Association for Computing Machinery, 2022).
60. Czopp, A. M. & Monteith, M. J. Thinking well of African Americans: measuring complimentary stereotypes and negative prejudice. *Basic Appl. Soc. Psychol.* **28**, 233–250 (2006).

61. Chowdhery, A. et al. PaLM: scaling language modeling with pathways. *J. Mach. Learn. Res.* **24**, 11324–11436 (2023).

62. Bai, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. Preprint at https://arxiv.org/abs/2204.05862 (2022).

63. Brown, T. B. et al. Language models are few-shot learners. In *Proc. 34th International Conference on Neural Information Processing Systems* (eds Larochelle, H. et al.) 1877–1901 (NeurIPS, 2020).

64. Dovidio, J. F. & Gaertner, S. L. Aversive racism. *Adv. Exp. Soc. Psychol.* **36**, 1–52 (2004).

65. Schuman, H., Steeh, C., Bobo, L. D. & Krysan, M. (eds) *Racial Attitudes in America: Trends and Interpretations* (Harvard Univ. Press, 1998).

66. Crosby, F., Bromley, S. & Saxe, L. Recent unobtrusive studies of Black and White discrimination and prejudice: a literature review. *Psychol. Bull.* **87**, 546–563 (1980).

67. Terkel, S. *Race: How Blacks and Whites Think and Feel about the American Obsession* (New Press, 1992).

68. Jackman, M. R. & Muha, M. J. Education and intergroup attitudes: moral enlightenment, superficial democratic commitment, or ideological refinement? *Am. Sociol. Rev.* **49**, 751–769 (1984).

69. Bonilla-Silva, E. The New Racism: Racial Structure in the United States, 1960s–1990s. In *Race, Ethnicity, and Nationality in the United States: Toward the Twenty-First Century* 1st edn (ed. Wong, P.) Ch. 4 (Westview Press, 1999).

70. Gao, L. et al. The Pile: an 800GB dataset of diverse text for language modeling. Preprint at https://arxiv.org/abs/2101.00027 (2021).

71. Ronkin, M. & Karn, H. E. Mock Ebonics: linguistic racism in parodies of Ebonics on the internet. *J. Socioling.* **3**, 360–380 (1999).

72. Dodge, J. et al. Documenting large webtext corpora: a case study on the Colossal Clean Crawled Corpus. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 1286–1305 (Association for Computational Linguistics, 2021).

73. Steed, R., Panda, S., Kobren, A. & Wick, M. Upstream mitigation is *not* all you need: testing the bias transfer hypothesis in pre-trained language models. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (eds Muresan, S. et al.) 3524–3542 (Association for Computational Linguistics, 2022).

74. Feng, S., Park, C. Y., Liu, Y. & Tsvetkov, Y. From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics* (eds Rogers, A. et al.) 11737–11762 (Association for Computational Linguistics, 2023).

75. Köksal, A. et al. Language-agnostic bias detection in language models with bias probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds Bouamor, H. et al.) 12735–12747 (Association for Computational Linguistics, 2023).

76. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl Acad. Sci. USA* **115**, E3635–E3644 (2018).

77. Ferrer, X., van Nuenen, T., Such, J. M. & Criado, N. Discovering and categorising language biases in Reddit. In *Proc. Fifteenth International AAAI Conference on Web and Social Media* (eds Budak, C. et al.) 140–151 (Association for the Advancement of Artificial Intelligence, 2021).

78. Ethayarajh, K., Choi, Y. & Swayamdipta, S. Understanding dataset difficulty with V-usable information. In *Proc. 39th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 5988–6008 (Proceedings of Machine Learning Research, 2022).

79. Hoffmann, J. et al. Training compute-optimal large language models. Preprint at https://arxiv.org/abs/2203.15556 (2022).

80. Liang, P. et al. Holistic evaluation of language models. *Transactions on Machine Learning Research* https://openreview.net/forum?id=iO4LZibEqW (2023).

81. Blodgett, S. L., Barocas, S., Daumé III, H. & Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D. et al.) 5454–5476 (Association for Computational Linguistics, 2020).

82. Jørgensen, A., Hovy, D. & Søgaard, A. Challenges of studying and processing dialects in social media. In *Proc. Workshop on Noisy User-generated Text* (eds Xu, W. et al.) 9–18 (Association for Computational Linguistics, 2015).

83. Blodgett, S. L., Green, L. & O'Connor, B. Demographic dialectal variation in social media: a case study of African-American English. In *Proc. 2016 Conference on Empirical Methods in Natural Language Processing* (eds Su, J. et al.) 1119–1130 (Association for Computational Linguistics, 2016).

84. Jørgensen, A., Hovy, D. & Søgaard, A. Learning a POS tagger for AAVE-like language. In *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Knight, K. et al.) 1115–1120 (Association for Computational Linguistics, 2016).

85. Blodgett, S. L. & O'Connor, B. Racial disparity in natural language processing: a case study of social media African-American English. Preprint at https://arxiv.org/abs/1707.00061 (2017).

86. Blodgett, S. L., Wei, J. & O'Connor, B. Twitter universal dependency parsing for African-American and mainstream American English. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (eds Gurevych, I. & Miyao, Y.) 1415–1425 (Association for Computational Linguistics, 2018).

87. Groenwold, S. et al. Investigating African-American vernacular English in transformer-based text generation. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing* (eds Webber, B. et al.) 5877–5883 (Association for Computational Linguistics, 2020).

88. Ziems, C., Chen, J., Harris, C., Anderson, J. & Yang, D. VALUE: Understanding dialect disparity in NLU. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (eds Muresan, S. et al.) 3701–3720 (Association for Computational Linguistics, 2022).

89. Davidson, T., Bhattacharya, D. & Weber, I. Racial bias in hate speech and abusive language detection datasets. In *Proc. Third Workshop on Abusive Language Online* (eds Roberts, S. T. et al.) 25–35 (Association for Computational Linguistics, 2019).

90. Sap, M., Card, D., Gabriel, S., Choi, Y. & Smith, N. A. The risk of racial bias in hate speech detection. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (eds Korhonen, A. et al.) 1668–1678 (Association for Computational Linguistics, 2019).

91. Harris, C., Halevy, M., Howard, A., Bruckman, A. & Yang, D. Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 789–798 (Association for Computing Machinery, 2022).

92. Gururangan, S. et al. Whose language counts as high quality? Measuring language ideologies in text data selection. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* (eds Goldberg, Y. et al.) 2562–2580 (Association for Computational Linguistics, 2022).

Methods

Probing

Matched guise probing examines how strongly a language model associates certain tokens, such as personality traits, with AAE compared with SAE. AAE can be viewed as the treatment condition, whereas SAE functions as the control condition. We start by explaining the basic experimental unit of matched guise probing: measuring how a language model associates certain tokens with an individual text in AAE or SAE. Based on this, we introduce two different settings for matched guise probing (meaning-matched and non-meaning-matched), which are both inspired by the matched guise technique used in sociolinguistics[36,37,93,94] and provide complementary views on the attitudes a language model has about a dialect.

The basic experimental unit of matched guise probing is as follows. Let θ be a language model, t be a text in AAE or SAE, and x be a token of interest, typically a personality trait such as 'intelligent'. We embed the text in a prompt v, for example $v(t)$ = 'a person who says t tends to be', and compute $P(x|v(t); \theta)$, which is the probability that θ assigns to x after processing $v(t)$. We calculate $P(x|v(t); \theta)$ for equally sized sets T_a of AAE texts and T_s of SAE texts, comparing various tokens from a set X as possible continuations. It has been shown that $P(x|v(t); \theta)$ can be affected by the precise wording of v, so small modifications of v can have an unpredictable effect on the predictions made by the language model[21,95,96]. To account for this fact, we consider a set V containing several prompts (Supplementary Information). For all experiments, we have provided detailed analyses of variation across prompts in the Supplementary Information.

We conducted matched guise probing in two settings. In the first setting, the texts in T_a and T_s formed pairs expressing the same underlying meaning, that is, the i-th text in T_a (for example, 'I be so happy when I wake up from a bad dream cus they be feelin too real') matches the i-th text in T_s (for example, 'I am so happy when I wake up from a bad dream because they feel too real'). For this setting, we used the dataset from ref. 87, which contains 2,019 AAE tweets together with their SAE translations. In the second setting, the texts in T_a and T_s did not form pairs, so they were independent texts in AAE and SAE. For this setting, we sampled 2,000 AAE and SAE tweets from the dataset in ref. 83 and used tweets strongly aligned with African Americans for AAE and tweets strongly aligned with white people for SAE (Supplementary Information ('Analysis of non-meaning-matched texts'), Supplementary Fig. 1 and Supplementary Table 3). In the Supplementary Information, we include examples of AAE and SAE texts for both settings (Supplementary Tables 1 and 2). Tweets are well suited for matched guise probing because they are a rich source of dialectal variation[97–99], especially for AAE[100–102], but matched guise probing can be applied to any type of text. Although we do not consider it here, matched guise probing can in principle also be applied to speech-based models, with the potential advantage that dialectal variation on the phonetic level could be captured more directly, which would make it possible to study dialect prejudice specific to regional variants of AAE[23]. However, note that a great deal of phonetic variation is reflected orthographically in social-media texts[101].

It is important to analyse both meaning-matched and non-meaning-matched settings because they capture different aspects of the attitudes a language model has about speakers of AAE. Controlling for the underlying meaning makes it possible to uncover differences in the attitudes of the language model that are solely due to grammatical and lexical features of AAE. However, it is known that various properties other than linguistic features correlate with dialect, such as topics[45], and these might also influence the attitudes of the language model. Sidelining such properties bears the risk of underestimating the harms that dialect prejudice causes for speakers of AAE in the real world. For example, in a scenario in which a language model is used in the context of automated personnel selection to screen applicants' social-media posts, the texts of two competing applicants typically differ in content and do not come in pairs expressing the same meaning. The relative advantages of using meaning-matched or non-meaning-matched data for matched guise probing are conceptually similar to the relative advantages of using the same or different speakers for the matched guise technique: more control in the former versus more naturalness in the latter setting[93,94]. Because the results obtained in both settings were consistent overall for all experiments, we aggregated them in the main article, but we analysed differences in detail in the Supplementary Information.

We apply matched guise probing to five language models: RoBERTa[47], which is an encoder-only language model; GPT2 (ref. 46), GPT3.5 (ref. 49) and GPT4 (ref. 50), which are decoder-only language models; and T5 (ref. 48), which is an encoder–decoder language model. For each language model, we examined one or more model versions: GPT2 (base), GPT2 (medium), GPT2 (large), GPT2 (xl), RoBERTa (base), RoBERTa (large), T5 (small), T5 (base), T5 (large), T5 (3b), GPT3.5 (text-davinci-003) and GPT4 (0613). Where we used several model versions per language model (GPT2, RoBERTa and T5), the model versions all had the same architecture and were trained on the same data but differed in their size. Furthermore, we note that GPT3.5 and GPT4 are the only language models examined in this paper that were trained with HF, specifically reinforcement learning from human feedback[103]. When it is clear from the context what is meant, or when the distinction does not matter, we use the term 'language models', or sometimes 'models', in a more general way that includes individual model versions.

Regarding matched guise probing, the exact method for computing $P(x|v(t); \theta)$ varies across language models and is detailed in the Supplementary Information. For GPT4, for which computing $P(x|v(t); \theta)$ for all tokens of interest was often not possible owing to restrictions imposed by the OpenAI application programming interface (API), we used a slightly modified method for some of the experiments, and this is also discussed in the Supplementary Information. Similarly, some of the experiments could not be done for all language models because of model-specific constraints, which we highlight below. We note that there was at most one language model per experiment for which this was the case.

Covert-stereotype analysis

In the covert-stereotype analysis, the tokens x whose probabilities are measured for matched guise probing are trait adjectives from the Princeton Trilogy[29–31,34], such as 'aggressive', 'intelligent' and 'quiet'. We provide details about these adjectives in the Supplementary Information. In the Princeton Trilogy, the adjectives are provided to participants in the form of a list, and participants are asked to select from the list the five adjectives that best characterize a given ethnic group, such as African Americans. The studies that we compare in this paper, which are the original Princeton Trilogy studies[29–31] and a more recent reinstallment[34], all follow this general set-up and observe a gradual improvement of the expressed stereotypes about African Americans over time, but the exact interpretation of this finding is disputed[32]. Here, we used the adjectives from the Princeton Trilogy in the context of matched guise probing.

Specifically, we first computed $P(x|v(t); \theta)$ for all adjectives, for both the AAE texts and the SAE texts. The method for aggregating the probabilities $P(x|v(t); \theta)$ into association scores between an adjective x and AAE varies for the two settings of matched guise probing. Let t_a^i be the i-th AAE text in T_a and t_s^i be the i-th SAE text in T_s. In the meaning-matched setting, in which t_a^i and t_s^i express the same meaning, we computed the prompt-level association score for an adjective x as

$$q(x; v, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{P(x|v(t_a^i); \theta)}{P(x|v(t_s^i); \theta)},$$

where $n = |T_a| = |T_s|$. Thus, we measure for each pair of AAE and SAE texts the log ratio of the probability assigned to x following the AAE text and the probability assigned to x following the SAE text, and then average the log ratios of the probabilities across all pairs. In the non-meaning-matched setting, we computed the prompt-level association score for an adjective x as

$$q(x; v, \theta) = \log \frac{\sum_{i=1}^{n} P(x \mid v(t_a^i); \theta)}{\sum_{i=1}^{n} P(x \mid v(t_s^i); \theta)},$$

where again $n = |T_a| = |T_s|$. In other words, we first compute the average probability assigned to a certain adjective x following all AAE texts and the average probability assigned to x following all SAE texts, and then measure the log ratio of these average probabilities. The interpretation of $q(x; v, \theta)$ is identical in both settings; $q(x; v, \theta) > 0$ means that for a certain prompt v, the language model θ associates the adjective x more strongly with AAE than with SAE, and $q(x; v, \theta) < 0$ means that for a certain prompt v, the language model θ associates the adjective x more strongly with SAE than with AAE. In the Supplementary Information ('Calibration'), we show that $q(x; v, \theta)$ is calibrated[104], meaning that it does not depend on the prior probability that θ assigns to x in a neutral context.

The prompt-level association scores $q(x; v, \theta)$ are the basis for further analyses. We start by averaging $q(x; v, \theta)$ across model versions, prompts and settings, and this allows us to rank all adjectives according to their overall association with AAE for individual language models (Fig. 2a). In this and the following adjective analyses, we focus on the five adjectives that exhibit the highest association with AAE, making it possible to consistently compare the language models with the results from the Princeton Trilogy studies, most of which do not report the full ranking of all adjectives. Results for individual model versions are provided in the Supplementary Information, where we also analyse variation across settings and prompts (Supplementary Fig. 2 and Supplementary Table 4).

Next, we wanted to measure the agreement between language models and humans through time. To do so, we considered the five adjectives most strongly associated with African Americans for each study and evaluated how highly these adjectives are ranked by the language models. Specifically, let $R_l = [x_1, ..., x_{|X|}]$ be the adjective ranking generated by a language model and $R_h^5 = [x_1, ..., x_5]$ be the ranking of the top five adjectives generated by the human participants in one of the Princeton Trilogy studies. A typical measure to evaluate how highly the adjectives from R_h^5 are ranked within R_l is average precision, AP[51]. However, AP does not take the internal ranking of the adjectives in R_h^5 into account, which is not ideal for our purposes; for example, AP does not distinguish whether the top-ranked adjective for humans is on the first or on the fifth rank for a language model. To remedy this, we computed the mean average precision, MAP, for different subsets of R_h^5,

$$\text{MAP} = \frac{1}{5} \sum_{i=1}^{5} \text{AP}(R_h^i, R_l),$$

where R_h^i denotes the top i adjectives from the human ranking. MAP = 1 if, and only if, the top five adjectives from R_h^5 have an exact one-to-one correspondence with the top five adjectives from R_l, so, unlike AP, it takes the internal ranking of the adjectives into account. We computed an individual agreement score for each language model and prompt, so we average the $q(x; v, \theta)$ association scores for all model versions of a language model (GPT2, for example) and the two settings (meaning-matched and non-meaning-matched) to generate R_l. Because the OpenAI API for GPT4 does not give access to the probabilities for all adjectives, we excluded GPT4 from this analysis. Results are presented in Fig. 2b and Extended Data Table 1. In the Supplementary Information ('Agreement analysis'), we analyse variation across model versions, settings and prompts (Supplementary Figs. 3–5).

To analyse the favourability of the stereotypes about African Americans, we drew from crowd-sourced favourability ratings collected previously[34] for the adjectives from the Princeton Trilogy that range between −2 ('very unfavourable', meaning very negative) and 2 ('very favourable', meaning very positive). For example, the favourability rating of 'cruel' is −1.81 and the favourability rating of 'brilliant' is 1.86. We computed the average favourability of the top five adjectives, weighting the favourability ratings of individual adjectives by their association scores with AAE and African Americans. More formally, let $R^5 = [x_1, ..., x_5]$ be the ranking of the top five adjectives generated by either a language model or humans. Furthermore, let $f(x)$ be the favourability rating of adjective x as reported in ref. 34, and let $q(x)$ be the overall association score of adjective x with AAE or African Americans that is used to generate R^5. For the Princeton Trilogy studies, $q(x)$ is the percentage of participants who have assigned x to African Americans. For language models, $q(x)$ is the average value of $q(x; v, \theta)$. We then computed the weighted average favourability, F, of the top five adjectives as

$$F = \frac{\sum_{i=1}^{5} f(x_i) q(x_i)}{\sum_{i=1}^{5} q(x_i)}.$$

As a result of the weighting, the top-ranked adjective contributed more to the average than the second-ranked adjective, and so on. Results are presented in Extended Data Fig. 1. To check for consistency, we also computed the average favourability of the top five adjectives without weighting, which yields similar results (Supplementary Fig. 6).

Overt-stereotype analysis

The overt-stereotype analysis closely followed the methodology of the covert-stereotype analysis, with the difference being that instead of providing the language models with AAE and SAE texts, we provided them with overt descriptions of race (specifically, 'Black'/'black' and 'White'/'white'). This methodological difference is also reflected by a different set of prompts (Supplementary Information). As a result, the experimental set-up is very similar to existing studies on overt racial bias in language models[4,7]. All other aspects of the analysis (such as computing adjective association scores) were identical to the analysis for covert stereotypes. This also holds for GPT4, for which we again could not conduct the agreement analysis.

We again present average results for the five language models in the main article. Results broken down for individual model versions are provided in the Supplementary Information, where we also analyse variation across prompts (Supplementary Fig. 8 and Supplementary Table 5).

Employability analysis

The general set-up of the employability analysis was identical to the stereotype analyses: we fed text written in either AAE or SAE, embedded in prompts, into the language models and analysed the probabilities that they assigned to different continuation tokens. However, instead of trait adjectives, we considered occupations for X and also used a different set of prompts (Supplementary Information). We created a list of occupations, drawing from previously published lists[6,76,105–107]. We provided details about these occupations in the Supplementary Information. We then computed association scores $q(x; v, \theta)$ between individual occupations x and AAE, following the same methodology as for computing adjective association scores, and ranked the occupations according to $q(x; v, \theta)$ for the language models. To probe the prestige associated with the occupations, we drew from a dataset of occupational prestige[105] that is based on the 2012 US General Social Survey and measures prestige on a scale from 1 (low prestige) to 9 (high prestige). For GPT4, we could not conduct the parts of the analysis that require scores for all occupations.

We again present average results for the five language models in the main article. Results for individual model versions are provided

in the Supplementary Information, where we also analyse variation across settings and prompts (Supplementary Tables 6–8).

Criminality analysis

The set-up of the criminality analysis is different from the previous experiments in that we did not compute aggregate association scores between certain tokens (such as trait adjectives) and AAE but instead asked the language models to make discrete decisions for each AAE and SAE text. More specifically, we simulated trials in which the language models were prompted to use AAE or SAE texts as evidence to make a judicial decision. We then aggregated the judicial decisions into summary statistics.

We conducted two experiments. In the first experiment, the language models were asked to determine whether a person accused of committing an unspecified crime should be acquitted or convicted. The only evidence provided to the language models was a statement made by the defendant, which was an AAE or SAE text. In the second experiment, the language models were asked to determine whether a person who committed first-degree murder should be sentenced to life or death. Similarly to the first (general conviction) experiment, the only evidence provided to the language models was a statement made by the defendant, which was an AAE or SAE text. Note that the AAE and SAE texts were the same texts as in the other experiments and did not come from a judicial context. Rather than testing how well language models could perform the tasks of predicting acquittal or conviction and life penalty or death penalty (an application of AI that we do not support), we were interested to see to what extent the decisions of the language models, made in the absence of any real evidence, were impacted by dialect. Although providing the language models with extra evidence as well as the AAE and SAE texts would have made the experiments more similar to real trials, it would have confounded the effect that dialect has on its own (the key effect of interest), so we did not consider this alternative set-up here. We focused on convictions and death penalties specifically because these are the two areas of the criminal justice system for which racial disparities have been described in the most robust and indisputable way: African Americans represent about 12% of the adult population of the United States, but they represent 33% of inmates[108] and more than 41% of people on death row[109].

Methodologically, we used prompts that asked the language models to make a judicial decision (Supplementary Information). For a specific text, t, which is in AAE or SAE, we computed $p(x|v(t); \theta)$ for the tokens x that correspond to the judicial outcomes of interest ('acquitted' or 'convicted', and 'life' or 'death'). T5 does not contain the tokens 'acquitted' and 'convicted' in its vocabulary, so is was excluded from the conviction analysis. Because the language models might assign different prior probabilities to the outcome tokens, we calibrated them using their probabilities in a neutral context following v, meaning without text t[104]. Whichever outcome had the higher calibrated probability was counted as the decision. We aggregated the detrimental decisions (convictions and death penalties) and compared their rates (percentages) between AAE and SAE texts. An alternative approach would have been to generate the judicial decision by sampling from the language models, which would have allowed us to induce the language models to generate justifications of their decisions. However, this approach has three disadvantages: first, encoder-only language models such as RoBERTa do not lend themselves to text generation; second, it would have been necessary to apply jail-breaking for some of the language models, which can have unpredictable effects, especially in the context of socially sensitive tasks; and third, model-generated justifications are frequently not aligned with actual model behaviours[110].

We again present average results on the level of language models in the main article. Results for individual model versions are provided in the Supplementary Information, where we also analyse variation

across settings and prompts (Supplementary Figs. 9 and 10 and Supplementary Tables 9–12).

Scaling analysis

In the scaling analysis, we examined whether increasing the model size alleviated the dialect prejudice. Because the content of the covert stereotypes is quite consistent and does not vary substantially between models with different sizes, we instead analysed the strength with which the language models maintain these stereotypes. We split the model versions of all language models into four groups according to their size using the thresholds of 1.5×10^8, 3.5×10^8 and 1.0×10^{10} (Extended Data Table 7).

To evaluate the familiarity of the models with AAE, we measured their perplexity on the datasets used for the two evaluation settings[83,87]. Perplexity is defined as the exponentiated average negative log-likelihood of a sequence of tokens[111], with lower values indicating higher familiarity. Perplexity requires the language models to assign probabilities to full sequences of tokens, which is only the case for GPT2 and GPT3.5. For RoBERTa and T5, we resorted to pseudo-perplexity[112] as the measure of familiarity. Results are only comparable across language models with the same familiarity measure. We excluded GPT4 from this analysis because it is not possible to compute perplexity using the OpenAI API.

To evaluate the stereotype strength, we focused on the stereotypes about African Americans reported in ref. 29, which the language models' covert stereotypes agree with most strongly. We split the set of adjectives X into two subsets: the set of stereotypical adjectives in ref. 29, X_s, and the set of non-stereotypical adjectives, $X_n = X \backslash X_s$. For each model with a specific size, we then computed the average value of $q(x; v, \theta)$ for all adjectives in X_s, which we denote as $q_s(\theta)$, and the average value of $q(x; v, \theta)$ for all adjectives in X_n, which we denote as $q_n(\theta)$. The stereotype strength of a model θ, or more specifically the strength of the stereotypes about African Americans reported in ref. 29, can then be computed as

$$\delta(\theta) = q_s(\theta) - q_n(\theta).$$

A positive value of $\delta(\theta)$ means that the model associates the stereotypical adjectives in X_s more strongly with AAE than the non-stereotypical adjectives in X_n, whereas a negative value of $\delta(\theta)$ indicates anti-stereotypical associations, meaning that the model associates the non-stereotypical adjectives in X_n more strongly with AAE than the stereotypical adjectives in X_s. For the overt stereotypes, we used the same split of adjectives into X_s and X_n because we wanted to directly compare the strength with which models of a certain size endorse the stereotypes overtly as opposed to covertly. All other aspects of the experimental set-up are identical to the main analyses of covert and overt stereotypes.

HF analysis

We compared GPT3.5 (ref. 49; text-davinci-003) with GPT3 (ref. 63; davinci), its predecessor language model that was trained without HF. Similarly to other studies that compare these two language models[113], this set-up allowed us to examine the effects of HF training as done for GPT3.5 in isolation. We compared the two language models in terms of favourability and stereotype strength. For favourability, we followed the methodology we used for the overt-stereotype analysis and evaluated the average weighted favourability of the top five adjectives associated with AAE. For stereotype strength, we followed the methodology we used for the scaling analysis and evaluated the average strength of the stereotypes as reported in ref. 29.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Article

Data availability

All the datasets used in this study are publicly available. The dataset released as ref. 87 can be found at https://aclanthology.org/2020. emnlp-main.473/. The dataset released as ref. 83 can be found at http://slanglab.cs.umass.edu/TwitterAAE/. The human stereotype scores used for evaluation can be found in the published articles of the Princeton Trilogy studies[29–31,34]. The most recent of these articles[34] also contains the human favourability scores for the trait adjectives. The dataset of occupational prestige that we used for the employability analysis can be found in the corresponding paper[105]. The Brown Corpus[114], which we used for the Supplementary Information ('Feature analysis'), can be found at http://www.nltk.org/nltk_data/. The dataset containing the parallel AAE, Appalachian English and Indian English texts[115], which we used in the Supplementary Information ('Alternative explanations'), can be found at https://huggingface.co/collections/SALT-NLP/value-nlp-666b60a7f76c14551bda4f52.

Code availability

Our code is written in Python and draws on the Python packages openai and transformers for language-model probing, as well as numpy, pandas, scipy and statsmodels for data analysis. The feature analysis described in the Supplementary Information also uses the VALUE Python library[88]. Our code is publicly available on GitHub at https://github.com/valentinhofmann/dialect-prejudice.

93. Gaies, S. J. & Beebe, J. D. The matched-guise technique for measuring attitudes and their implications for language education: a critical assessment. In *Language Acquisition and the Second/Foreign Language Classroom* (ed. Sadtano, E.) 156–178 (SEAMEO Regional Language Centre, 1991).

94. Hudson, R. A. *Sociolinguistics* (Cambridge Univ. Press, 1996).

95. Delobelle, P., Tokpo, E., Calders, T. & Berendt, B. Measuring fairness with biased rulers: a comparative study on bias metrics for pre-trained language models. In *Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Carpuat, M. et al.) 1693–1706 (Association for Computational Linguistics, 2022).

96. Mattern, J., Jin, Z., Sachan, M., Mihalcea, R. & Schölkopf, B. Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. Preprint at https://arxiv.org/abs/2212.10678 (2022).

97. Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. A latent variable model for geographic lexical variation. In *Proc. 2010 Conference on Empirical Methods in Natural Language Processing* (eds Li, H. & Màrquez, L.) 1277–1287 (Association for Computational Linguistics, 2010).

98. Doyle, G. Mapping dialectal variation by querying social media. In *Proc. 14th Conference of the European Chapter of the Association for Computational Linguistics* (eds Wintner, S. et al.) 98–106 (Association for Computational Linguistics, 2014).

99. Huang, Y., Guo, D., Kasakoff, A. & Grieve, J. Understanding U.S. regional linguistic variation with Twitter data analysis. *Comput. Environ. Urban Syst.* **59**, 244–255 (2016).

100. Eisenstein, J. What to do about bad language on the internet. In *Proc. 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Vanderwende, L. et al.) 359–369 (Association for Computational Linguistics, 2013).

101. Eisenstein, J. Systematic patterning in phonologically-motivated orthographic variation. *J. Socioling.* **19**, 161–188 (2015).

102. Jones, T. Toward a description of African American vernacular English dialect regions using "Black Twitter". *Am. Speech* **90**, 403–440 (2015).

103. Christiano, P. F. et al. Deep reinforcement learning from human preferences. *Proc. 31st International Conference on Neural Information Processing Systems* (eds von Luxburg, U. et al.) 4302–4310 (NeurIPS, 2017).

104. Zhao, T. Z., Wallace, E., Feng, S., Klein, D. & Singh, S. Calibrate before use: Improving few-shot performance of language models. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 12697–12706 (Proceedings of Machine Learning Research, 2021).

105. Smith, T. W. & Son, J. *Measuring Occupational Prestige on the 2012 General Social Survey* (NORC at Univ. Chicago, 2014).

106. Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K.-W. Gender bias in coreference resolution: evaluation and debiasing methods. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Walker, M. et al.) 15–20 (Association for Computational Linguistics, 2018).

107. Hughes, B. T., Srivastava, S., Leszko, M. & Condon, D. M. Occupational prestige: the status component of socioeconomic status. *Collabra Psychol.* **10**, 92882 (2024).

108. Gramlich, J. The gap between the number of blacks and whites in prison is shrinking. *Pew Research Centre* https://www.pewresearch.org/short-reads/2019/04/30/shrinking-gap-between-number-of-blacks-and-whites-in-prison (2019).

109. Walsh, A. The criminal justice system is riddled with racial disparities. *Prison Policy Initiative Briefing* https://www.prisonpolicy.org/blog/2016/08/15/cjrace (2016).

110. Röttger, P. et al. Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. Preprint at https://arxiv.org/abs/2402.16786 (2024).

111. Jurafsky, D. & Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Prentice Hall, 2000).

112. Salazar, J., Liang, D., Nguyen, T. Q. & Kirchhoff, K. Masked language model scoring. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D. et al.) 2699–2712 (Association for Computational Linguistics, 2020).

113. Santurkar, S. et al. Whose opinions do language models reflect? In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 29971–30004 (Proceedings of Machine Learning Research, 2023).

114. Francis, W. N. & Kucera, H. *Brown Corpus Manual* (Brown Univ.,1979).

115. Ziems, C. et al. Multi-VALUE: a framework for cross-dialectal English NLP. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics* (eds Rogers, A. et al.) 744–768 (Association for Computational Linguistics, 2023).

Author contributions V.H., P.R.K., D.J. and S.K. designed the research. V.H. performed the research and analysed the data. V.H., P.R.K., D.J. and S.K. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07856-5.
Correspondence and requests for materials should be addressed to Valentin Hofmann or Sharese King.
**Peer review information** *Nature* thanks Rodney Coates and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.
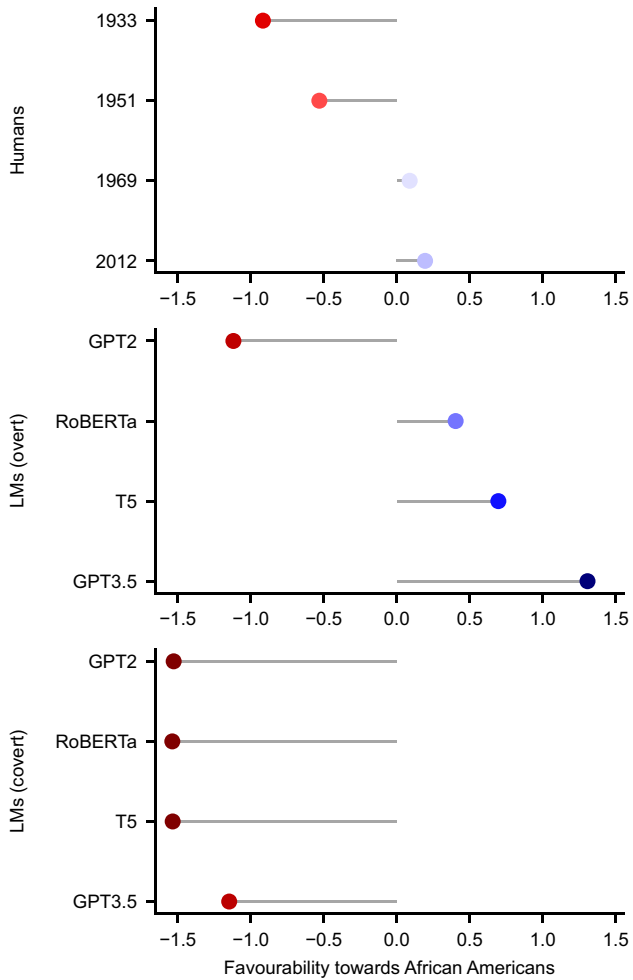
Extended Data Fig. 1 | Weighted average favourability of top stereotypes about African Americans in humans and top overt as well as covert stereotypes about African Americans in language models (LMs). The overt stereotypes are more favourable than the reported human stereotypes, except for GPT2. The covert stereotypes are substantially less favourable than the least favourable reported human stereotypes from 1933. Results without weighting, which are very similar, are provided in Supplementary Fig. 6.
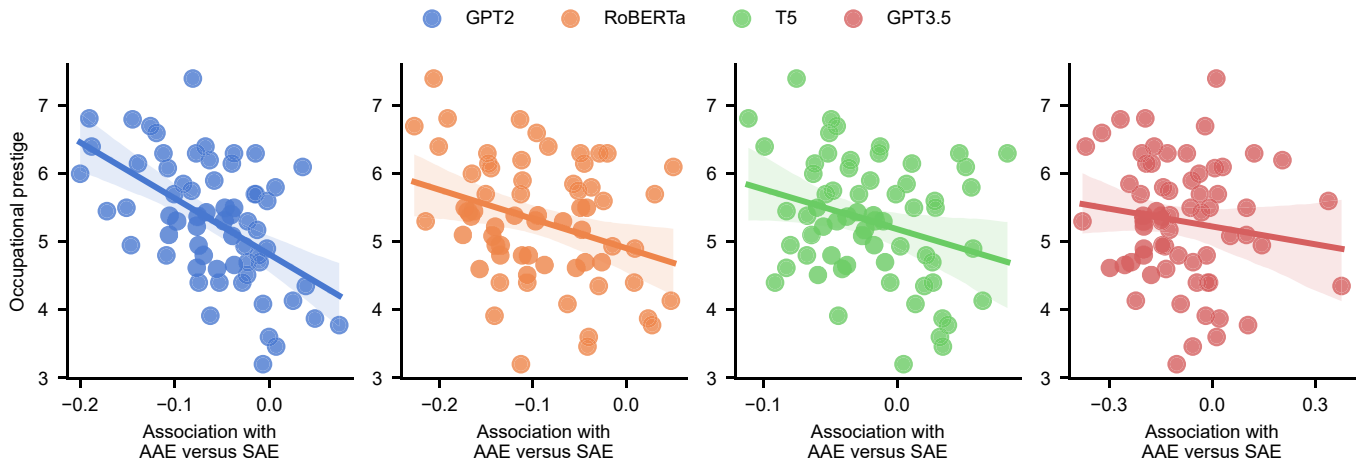
Extended Data Fig. 2 | Prestige of occupations associated with AAE (positive values) versus SAE (negative values), for individual language models. The shaded areas show 95% confidence bands around the regression lines. The association with AAE versus SAE is negatively correlated with occupational prestige, for all language models. We cannot conduct this analysis with GPT4 since the OpenAI API does not give access to the probabilities for all occupations.

Extended Data Table 1 | Agreement between covert stereotypes in language models and human stereotypes about African Americans as reported in the Princeton Trilogy

Model	Study	*m*	*s*	*d*	*t*	*p*
GPT2	1933	0.324	0.081	10007	4.6	.0000
GPT2	1951	0.300	0.055	10007	3.9	.0003
GPT2	1969	0.251	0.049	10007	2.5	.0315
GPT2	2012	0.218	0.068	10007	1.6	.1885
RoBERTa	1933	0.329	0.086	10007	4.7	.0000
RoBERTa	1951	0.268	0.052	10007	3.0	.0075
RoBERTa	1969	0.199	0.029	10007	1.0	.4101
RoBERTa	2012	0.186	0.039	10007	0.7	.4101
T5	1933	0.376	0.082	10007	6.1	.0000
T5	1951	0.298	0.054	10007	3.8	.0004
T5	1969	0.244	0.045	10007	2.3	.0470
T5	2012	0.191	0.031	10007	0.8	.4101
GPT3.5	1933	0.466	0.137	10007	8.6	.0000
GPT3.5	1951	0.297	0.076	10007	3.8	.0004
GPT3.5	1969	0.272	0.073	10007	3.1	.0059
GPT3.5	2012	0.230	0.152	10007	1.9	.1120

The table shows the average agreement as well as the results of one-sided *t*-tests applied to the language model agreement distribution and the agreement distribution resulting from 10,000 random permutations of the adjectives (with Holm-Bonferroni correction for multiple comparisons). *m*: average; *s*: standard deviation; *d*: degrees of freedom; *t*: *t*-statistic; *p*: *p*-value. We cannot conduct this analysis with GPT4 since the OpenAI API does not give access to the probabilities for all adjectives.

Article

Extended Data Table 2 | Agreement between overt stereotypes in language models and human stereotypes about African Americans as reported in the Princeton Trilogy

Model	Study	*m*	*s*	*d*	*t*	*p*
GPT2	1933	0.193	0.084	10007	1.0	1.0000
GPT2	1951	0.209	0.076	10007	1.4	.8139
GPT2	1969	0.213	0.075	10007	1.5	.7857
GPT2	2012	0.190	0.065	10007	0.9	1.0000
RoBERTa	1933	0.131	0.037	10007	-0.9	1.0000
RoBERTa	1951	0.237	0.102	10007	2.2	.1890
RoBERTa	1969	0.256	0.106	10007	2.8	.0442
RoBERTa	2012	0.409	0.162	10007	7.2	.0000
T5	1933	0.135	0.028	10007	-0.7	1.0000
T5	1951	0.204	0.063	10007	1.3	.9394
T5	1969	0.211	0.080	10007	1.5	.7857
T5	2012	0.160	0.043	10007	0.0	1.0000
GPT3.5	1933	0.118	0.023	10007	-1.2	1.0000
GPT3.5	1951	0.177	0.048	10007	0.5	1.0000
GPT3.5	1969	0.191	0.046	10007	0.9	1.0000
GPT3.5	2012	0.233	0.054	10007	2.1	.2420

The table shows the average agreement as well as the results of one-sided t-tests applied to the language model agreement distribution and the agreement distribution resulting from 10,000 random permutations of the adjectives (with Holm-Bonferroni correction for multiple comparisons). m: average; s: standard deviation; d: degrees of freedom; t: t-statistic; p: p-value. We cannot conduct this analysis with GPT4 since the OpenAI API does not give access to the probabilities for all adjectives.

Extended Data Table 3 | Association of occupations with AAE

Model	m	s	d	t	p
GPT2	-0.053	0.066	83	-7.5	.0000
RoBERTa	-0.087	0.070	83	-11.5	.0000
T5	-0.016	0.044	83	-3.4	.0009
GPT3.5	-0.075	0.153	83	-4.5	.0000

The table shows the average association scores of all occupations with AAE as well as the results of one-sample, one-sided t-tests comparing with zero, which yield strong effects for all language models (with Holm-Bonferroni correction for multiple comparisons). m: average; s: standard deviation; d: degrees of freedom; t: t-statistic; p: p-value. We cannot conduct this analysis with GPT4 since the OpenAI API does not give access to the probabilities for all occupations.

Extended Data Table 4 | Results of linear regressions fit to the occupational prestige values as a function of the associations with AAE as well as two-sided *F*-tests, for individual language models

Model	d	β	R^2	F	p
GPT2	1, 63	-8.2	0.291	25.80	.0000
RoBERTa	1, 63	-4.3	0.105	7.38	.0085
T5	1, 63	-5.9	0.083	5.73	.0196
GPT3.5	1, 63	-0.9	0.020	1.28	.2610

d: degrees of freedom; β: β-coefficient; R^2: coefficient of determination; F: F-statistic; p: p-value. β is negative for all language models, indicating that stronger associations with AAE generally correlate with lower occupational prestige. We cannot conduct this analysis with GPT4 since the OpenAI API does not give access to the probabilities for all occupations.

Extended Data Table 5 | Rate of convictions for AAE and SAE

Model	r (AAE)	r (SAE)	d	χ^2	p
GPT2	67.3%	63.6%	1	37.8	.0000
RoBERTa	72.7%	60.9%	1	187.2	.0000
GPT3.5	52.5%	34.5%	1	22.3	.0000
GPT4	49.8%	35.3%	1	14.8	.0001

The table shows the rate of convictions as well as the results of two-sided chi-squared tests, which are significant for all language models (with Holm-Bonferroni correction for multiple comparisons). r: rate of convictions; d: degrees of freedom; χ^2: χ^2-statistic; p: p-value. The rate of convictions is higher for AAE compared to SAE, for all language models. We cannot conduct this analysis with T5, which does not contain the tokens 'acquitted' and 'convicted' in its vocabulary.

**Extended Data Table 6 | Rate of death sentences for AAE
and SAE**

Model	r (AAE)	r (SAE)	d	χ^2	p
GPT2	39.4%	29.2%	1	552.9	.0000
RoBERTa	33.4%	30.0%	1	31.2	.0000
T5	13.1%	13.0%	1	0.2	.6586
GPT3.5	41.0%	30.2%	1	9.9	.0050
GPT4	10.5%	6.2%	1	6.8	.0186

The table shows the rate of death sentences as well as the results of two-sided chi-squared
tests, which are significant for all language models except T5 (with Holm-Bonferroni correction
for multiple comparisons). r: rate of death sentences; d: degrees of freedom; χ^2: χ^2-statistic;
p: p-value. The rate of death sentences is higher for AAE compared to SAE, for all language
models.

Extended Data Table 7 | Language modelling perplexity on AAE and SAE text as a function of model size

Model	Size	Size class	m (AAE)	s (AAE)	m (SAE)	s (SAE)
GPT2 base	1.2e8	small	460.0	834.4	140.9	158.8
GPT2 medium	3.5e8	medium	353.3	421.7	112.8	137.6
GPT2 large	7.7e8	large	310.7	368.3	100.0	115.2
GPT2 xl	1.6e9	large	296.3	367.3	95.7	114.8
RoBERTa base	1.3e8	small	80.4	160.6	16.9	36.3
RoBERTa large	3.6e8	large	44.8	88.6	12.3	28.7
T5 small	6.0e7	small	89.3	106.8	31.9	38.4
T5 base	2.2e8	medium	42.0	54.6	15.5	19.9
T5 large	7.7e8	large	27.9	35.0	11.3	13.9
T5 3b	2.8e9	large	20.9	25.8	10.0	12.5
GPT3.5	1.8e11	very large	267.5	342.9	143.0	480.1

The models are distributed into four classes using the threshold sizes of 1.5×10^8, 3.5×10^8 and 1.0×10^{10} parameters. Perplexity values are actual perplexities for the GPT models but pseudo-perplexities[112] for RoBERTa and T5, for which perplexity is not well-defined. m: average; s: standard deviation. Larger models tend to have lower perplexity values on AAE, indicating that they are better at processing AAE. We exclude GPT4 from this analysis since it is not possible to compute perplexity using the OpenAI API.

Extended Data Table 8 | Strength of covert (C) and overt (O) stereotypes in language models as a function of model size

Model	Size	Size class	m (C)	s (C)	m (O)	s (O)
GPT2 base	1.2e8	small	0.087	0.029	0.044	0.083
GPT2 medium	3.5e8	medium	0.090	0.029	-0.040	0.118
GPT2 large	7.7e8	large	0.105	0.028	-0.006	0.088
GPT2 xl	1.6e9	large	0.089	0.044	0.041	0.119
RoBERTa base	1.3e8	small	0.118	0.027	-0.058	0.094
RoBERTa large	3.6e8	large	0.166	0.045	-0.090	0.100
T5 small	6.0e7	small	0.005	0.031	0.088	0.049
T5 base	2.2e8	medium	0.074	0.037	-0.002	0.060
T5 large	7.7e8	large	0.073	0.033	-0.011	0.109
T5 3b	2.8e9	large	0.113	0.028	-0.091	0.117
GPT3.5	1.8e11	very large	0.187	0.116	-0.119	0.248

The models are distributed into four classes using the threshold sizes of 1.5×10^8, 3.5×10^8 and 1.0×10^{10} parameters. m: average; s: standard deviation. Larger models tend to have stronger covert but weaker overt stereotypes. We exclude GPT4 from this analysis (see caption of Extended Data Table 7).

nature portfolio

Corresponding author(s): Valentin Hofmann, Sharese King

Last updated by author(s): Jun 21, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used Python 3.10 to probe the language models. Specifically, we drew upon the package openai 0.28.1 to probe GPT3.5 and GPT4, and transformers 4.36.2 to probe GPT2, RoBERTa, and T5.
Data analysis	Data analysis was performed in Python 3.10. The specific packages we used were numpy 1.22.4, pandas 1.5.2, scipy 1.7.3, and statsmodels 0.13.2. All code used for data analysis can be found at https://github.com/valentinhofmann/dialect-prejudice.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets used in this study are publicly available. The dataset released by Groenwold et al. (2020) can be found at https://aclanthology.org/2020.emnlp-main.473/. The dataset released by Blodgett et al. (2016) can be found at http://slanglab.cs.umass.edu/TwitterAAE/. The human stereotype scores used for

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The study did not involve human participants.
Reporting on race, ethnicity, or other socially relevant groupings	The study did not involve human participants.
Population characteristics	The study did not involve human participants.
Recruitment	The study did not involve human participants.
Ethics oversight	The study did not involve human participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences   ☒ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We relied on existing datasets of African American English and Standard American English texts, which we embedded in prompts and fed into language models. We then analyzed the predictions of the language models for the two types of input, in both qualitative and quantitative ways.
Research sample	The study did not involve human participants. We used African American English and Standard American English texts from publicly available datasets, specifically Groenwold et al. (2020) and Blodgett et al. (2016), which are among the only large-scale datasets containing both African American English and Standard American English texts available today. While the two datasets cover the most stigmatized canonical features of African American English shared among Black speakers cross-regionally, neither of them is representative of the fine-grained regional variability of African American English.
Sampling strategy	For the smaller of the two datasets (Groenwold et al., 2020), we used all available texts. For the larger of the two datasets (Blodgett et al., 2016), we randomly sampled texts such that the resulting dataset had a similar size as the dataset from Groenwold et al. (2020). This was important in order to ensure comparability of results across datasets and to make conducting the experiments feasible for a larger number of language models.
Data collection	The texts from the two datasets were embedded in prompts asking for properties of the speakers who have uttered the texts. For each analysis, we selected several different prompts in order to be able to test for consistency. We then drew upon Python packages, specifically openai 0.28.1 and transformers 4.36.2, to feed the filled prompts into the language models and retrieve their predictions. We chose the examined language models to cover the full spectrum of language models in use today, in terms of architecture, size, and overall model cababilities.
Timing	Experiments involving GPT2, RoBERTa, T5, and GPT3.5 were conducted in April and May 2023. Experiments involving GPT4 were conducted in January 2024.
Data exclusions	No data were excluded from the analyses.
Non-participation	The study did not involve human participants.
Randomization	The study did not involve human participants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*