scientific reports

Check for updates

OPEN

AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably

Brian Porter✉ & Edouard Machery

As AI-generated text continues to evolve, distinguishing it from human-authored content has become increasingly difficult. This study examined whether non-expert readers could reliably differentiate between AI-generated poems and those written by well-known human poets. We conducted two experiments with non-expert poetry readers and found that participants performed below chance levels in identifying AI-generated poems (46.6% accuracy, $\chi^2(1, N = 16{,}340) = 75.13$, $p < 0.0001$). Notably, participants were more likely to judge AI-generated poems as human-authored than actual human-authored poems ($\chi^2(2, N = 16{,}340) = 247.04$, $p < 0.0001$). We found that AI-generated poems were rated more favorably in qualities such as rhythm and beauty, and that this contributed to their mistaken identification as human-authored. Our findings suggest that participants employed shared yet flawed heuristics to differentiate AI from human poetry: the simplicity of AI-generated poems may be easier for non-experts to understand, leading them to prefer AI-generated poetry and misinterpret the complexity of human poems as incoherence generated by AI.

Keywords Artificial intelligence, Generative AI, Poetry, Large language model, Human-AI interaction

Perception and preference in poetry: biases toward AI-generated poems

AI-generated images have become indistinguishable from reality. AI-generated paintings are judged to be human-created artworks at higher rates than actual human-created paintings[1]; AI-generated faces are judged to be real human faces at higher rate than actual photos of human faces[2–5], and AI-generated humor is just as funny as human-generated jokes[6]. Despite this, studies have consistently found a bias against AI-generated artwork; when told that an artwork is AI-generated, participants rate the work as lower quality[2,7].

Meanwhile, generative language algorithms have made significant progress towards human-level performance. Large language models (LLMs) like OpenAI's GPT-3[8] and Meta's Llama 2[9] have been trained on millions of tokens, and can produce texts that closely resemble human-written text. Some kinds of AI-generated text are already indistinguishable from human-written texts[10,11].

However, it has been argued that LLMs will not be able to generate high quality poetry, even if they reach human-level competence at other forms of text, because poetry depends on creativity and meaning, while AI-generated text is inherently uncreative and meaningless[12]. Poetry is a particularly difficult literary genre to understand and interpret, especially for non-experts; it "incorporates a degree of arbitrariness since there are no strict or universal rules for what is acceptable or not" and it "not only resists commonly acceptable meaning, but also reverses it"[13]. However, there has been great success in poetry generation in the field of computational creativity; Linardaki[13] provides a survey and discussion of work in poetry generation. By many metrics, specialized AI models are able to produce high-quality poetry.

Despite this success, evidence about non-experts' ability to distinguish AI-generated poetry has been mixed. Non-experts in poetry may use different cues, and be less familiar with the structural requirements of rhyme and meter, than experts in poetry or poetry generation. Gunser and colleagues[14] and Rahmeh[15] find that human-written poems are evaluated more positively than AI-generated poems. Köbis and Mossink[16] finds that when a human chooses the best AI-generated poem ("human-in-the-loop") participants cannot distinguish AI-generated poems from human-written poems, but when an AI-generated poem is chosen at random ("human-out-of-

Department of History and Philosophy of Science, University of Pittsburgh, Cathedral of Learning, Pittsburgh, PA 15260, USA. ✉email: brian.porter@pitt.edu

the-loop"), participants are able to distinguish AI-generated from human-written poems. They also find that participants evaluate AI-generated poems more negatively than human-written poems, regardless of whether or not participants are told that the poems were generated by AI. Hitsuwari et al.[17] finds that haikus created by AI with human intervention ("human-in-the-loop") are rated more highly than human-generated haikus or haikus generated by AI without human intervention ("human-out-of-the-loop"); they found no difference in ratings between human-written haikus and haikus generated by AI without human intervention.

Here, we extend prior work by showing that AI-generated poetry has reached the level of AI-generated images in non-expert assessments: across multiple eras and genres of poetry, non-expert participants cannot distinguish human-written poetry from poems generated by AI without human intervention or specialized fine-tuning. Like AI-generated paintings and faces, AI-generated poems are now "more human than human": we find that participants are more likely to judge that AI-generated poems are human-authored, compared to actual human-authored poems. Contrary to previous studies, we also find that participants rate AI-generated poems more highly than human-written poems across several qualitative dimensions. However, we confirm earlier findings that participants evaluate poems more negatively when *told* that the poem is generated by AI, as opposed to being told the poem is human-written.

We use these findings to offer a partial explanation of the "more human than human" phenomenon: non-expert poetry readers prefer the more accessible AI-generated poetry, which communicate emotions, ideas, and themes in more direct and easy-to-understand language, but expect AI-generated poetry to be worse; they therefore mistakenly interpret their own preference for a poem as evidence that it is human-written.

To summarize, we set out to determine (1) whether people can distinguish AI-generated poems from professional human-written poems, (2) what features of a poem people use to make those judgments, (3) whether perceptions of poems as human-written or AI-generated affect qualitative assessments of the poems, and (4) whether the actual authorship of a poem affects qualitative assessments of the poems.

To investigate these questions, we conducted 2 experiments. We collected 5 poems each from 10 well-known English-language poets, spanning much of the history of English poetry: Geoffrey Chaucer (1340s-1400), William Shakespeare (1564-1616), Samuel Butler (1613-1680), Lord Byron (1788-1824), Walt Whitman (1819-1892), Emily Dickinson (1830-1886), T.S. Eliot (1888-1965), Allen Ginsberg (1926-1997), Sylvia Plath (1932-1963), and Dorothea Lasky (1978-). Using ChatGPT 3.5, we generated 5 poems "in the style of" each poet. We used a "human out of the loop" paradigm[16]: we used the first 5 poems generated, and did not select the "best" out of a group of poems or provide any feedback or instructions to the model beyond "Write a short poem in the style of <poet> ". In the first experiment, 1,634 participants were randomly assigned to one of the 10 poets, and presented with 10 poems in random order: 5 poems written by that poet, and 5 generated by AI "in the style of" that poet. For each poem, participants were asked whether they thought the poem was generated by AI or written by a human poet.

To investigate how participants perceived and assessed AI-generated poetry, we conducted a second experiment: a qualitative assessment task. We recruited a new sample of 696 participants from Prolific. We used a randomly selected subset of the original 100 poems (10 poems total, one from each poet, 5 real and 5 AI-generated), and asked participants to assess each poem along 14 qualitative dimensions. Participants were randomly assigned to one of three framing conditions: "told human", in which participants were told that all poems were written by the professional human poet, regardless of actual authorship; "told AI", in which participants were told that all poems were generated by AI, regardless of actual authorship; and "told nothing", in which participants were not told anything about the poem's authorship. Participants in the "told nothing" condition were asked, after assessing each poem, whether they thought the poem was written by a human poet or generated by AI.

Results

Study 1: distinguishing AI-generated from human-written poems

As specified in our pre-registration (https://osf.io/5j4w9), we predicted that participants would be at chance when trying to identify AI-generated vs. human-written poems, setting the significance level at 0.005[18]; p's between 0.05 and 0.005 are "suggestive". Observed accuracy was in fact slightly lower than chance (46.6%, $\chi^2(1, N = 16340) = 75.13$, $p < 0.0001$). Observed agreement between participants was poor, but was higher than chance (Fleiss's kappa = 0.005, $p < 0.001$). Poor agreement suggests that, as expected, participants found the task very difficult, and were at least in part answering randomly. However, as in[10], the below-chance performance and the significant agreement between participants led us to conclude that participants were not answering *entirely* at random; they must be using at least some shared, yet mistaken, heuristics to distinguish AI-generated poems from human-written poems.

Participants were more likely to guess that AI-generated poems were written by humans than they were for actual human-written poems ($\chi^2(2, N = 16340) = 247.04$, w = 0.123, $p < 0.0001$). The five poems with the lowest rates of "human" ratings were all written by actual human poets; four of the five poems with the highest rates of "human" ratings were generated by AI.

We used a general linear mixed model logistic regression analysis (fit to a binomial distribution) to predict participant responses ("written by a human" or "generated by AI") with poem's authorship (human or AI), the identity of the poet, and their interaction as fixed effects. We used a sum coding for the identity of the poet, to interpret more easily the main effect of authorship across poets. As specified in our pre-registration, we initially included three random effects: random intercepts for participants (since we took 10 repeated measurements, one per poem, for each participant), random intercepts for poems, and random slopes for the identity of the poet for each poem. Following[19], we used principal component analysis (PCA) to check for overparameterization, and determined that the model was indeed overparameterized. PCA indicated that the random intercept for participants and the random slope for the identity of the poet were unnecessary and were causing the

overparameterization. This conclusion is borne out in the data; looking at the proportion of "written by a human" responses for each participant, the variance is only 0.021; the variance between poets is only 0.00013. The lower-than-expected variance in the data simply does not support the complex random-effects structure. We therefore fit a reduced model with random intercepts for poems as the only random effect. Using ANOVA to compare model fit, we found that the full model containing our original set of random effects (npar = 76, AIC = 22385, BIC = 22970, logLik = -11116) did not provide a significantly better fit than the reduced model (npar = 21, AIC = 22292.5, BIC = 22454.2, logLik = -11125.2). We therefore proceed with the reduced model.

The total explanatory power of the model was low (Conditional $R^2 = 0.024$, Marginal $R^2 = 0.013$), reflecting the expected difficulty of the discrimination task and the fact that, as a result, participants' answers differed only slightly from chance. Consistent with the deviation from chance in overall accuracy, authorship was significantly predictive of participant responses (b = -0.27716, SE = 0.04889, z = -5.669, $p < 0.0001$): being written by a human poet *decreased* the likelihood that a participant would respond that the poem was written by a human poet. The odds that a human-written poem is judged to be human-written are roughly 75% that of an AI-generated poem being judged human-authored (OR = 0.758). Full results can be found in our supplementary materials.

As an exploratory analysis, we refit the model with the addition of several variables reflecting structural features of the stimuli. Following[10], which found that participants use flawed heuristics based on grammar and vocabulary cues to identify AI-generated texts, we examined whether participants look to structural and grammatical features of the poems to determine authorship. To test this, we added to the previous model stimulus word count (scaled), stimulus line count (scaled), stimulus all-lines-rhyme (a binary variable indicating whether or not all lines in the poem end with a rhyme), stimulus quatrain (a binary variable indicating whether the poem was formatted entirely in four-line stanzas, i.e., "quatrains"), and stimulus first person (a variable reflecting whether or not the poem was written in first person, with 3 values: "I" if written in singular first person, "we" if written in plural first person, and "no" if not written in first person).

As expected, the total explanatory power of the model was low (Conditional $R^2 = 0.0024$, Marginal $R^2 = 0.017$). None of the structural features were significantly predictive, but both stimulus line count (b = 0.1461249, SE = 0.0661922, z = 2.208, $p = 0.02727$) and stimulus all-lines-rhyme (b = 0.2084246, SE = 0.0861658, z = 2.419, p = 0.01557) were suggestive. The effect of authorship (b = -0.1852979, SE = 0.0914278, z = -2.027, $p = 0.04269$) also appears to be somewhat weakened by the poem structural features; controlling for the structural features, the estimated odds of a human-authored poem being judged human-authored are roughly 83% that of an AI-generated poem (OR = 0.831). This suggests that participants are using some shared heuristics to discriminate AI-generated poems from human-written poems; they may take AI to be less able to form rhymes, and less able to produce longer poems. If so, these heuristics are flawed; in our dataset, AI-generated poems are in fact *more* likely to rhyme at all lines: 89% of our AI-generated poems rhyme, while only 40% of our human-written poems rhyme. There is also no significant difference in average number of lines between AI-generated poems and human-written poems in our dataset.

The effect of experience with poetry

We asked participants several questions to gauge their experience with poetry, including how much they like poetry, how frequently they read poetry, and their level of familiarity with their assigned poet. Overall, our participants reported a low level of experience with poetry: 90.4% of participants reported that they read poetry a few times per year or less, 55.8% described themselves as "not very familiar with poetry", and 66.8% describe themselves as "not familiar at all" with their assigned poet. Full details of the participant responses to these questions can be found in table S1 in our supplementary materials.

In order to determine if experience with poetry improves discrimination accuracy, we ran an exploratory model using variables for participants' answers to our poetry background and demographics questions. We included self-reported confidence, familiarity with the assigned poet, background in poetry, frequency of reading poetry, how much participants like poetry, whether or not they had ever taken a poetry course, age, gender, education level, and whether or not they had seen any of the poems before. Confidence was scaled, and we treated poet familiarity, poetry background, read frequency, liking poetry, and education level as ordered factors. We used this model to predict not whether participants answered "AI" or "human," but whether participants answered the question correctly (e.g., answered "generated by AI" when the poem was actually generated by AI). As specified in our pre-registration, we predicted that participant expertise or familiarity with poetry would make no difference in discrimination performance. This was largely confirmed; the explanatory power of the model was low (McFadden's $R^2 = 0.012$), and none of the effects measuring poetry experience had a significant positive effect on accuracy. Confidence had a small but significant negative effect (b = -0.021673, SE = 0.003986, z = -5.437, $p < 0.0001$), indicating that participants were slightly more likely to guess incorrectly when they were more confident in their answer.

We find two positive effects on discrimination accuracy: gender, specifically "non-binary/third gender" (b = 0.169080, SE = 0.030607, z = 5.524, $p < 0.0001$), and having seen any of the poems before (b = 0.060356, SE = 0.016726, z = 3.608, $p = 0.000309$). These effects are very small; having seen poems before only increases the odds of a correct answer by 6% (OR = 1.062). These findings suggest that experience with poetry did not improve discrimination performance unless that experience allowed them to recognize the specific poems used in the study. In summary, Study 1 showed that human-out-of-the-loop AI-generated poetry is judged to be human-written more often than poetry written by actual human poets, and that experience with poetry does not improve discrimination performance. Our results contrast with those of previous studies, in which participants were able to distinguish the poems of professional poets from human-out-of-the-loop AI-generated poetry[16], or that participants are at chance in distinguishing human poetry from human-out-of-the-loop AI-generated poetry[17]. Past research has suggested that AI-generated poetry needs human intervention to seem human-

written to non-expert participants, but recent advances in LLMs have achieved a new state-of-the-art in human-out-of-the-loop AI poetry that now, to our participants, seems "more human than human."

Study 2: evaluating AI-generated and human-generated poems

Our second study asks participants to rate each poem's overall quality, rhythm, imagery, sound; the extent to which the poem was moving, profound, witty, lyrical, inspiring, beautiful, meaningful, and original; and how well the poem conveyed a specific theme, and how well it conveyed a specific mood or emotion. Each of these was reported on a 7-point Likert scale. In addition to these 14 qualitative assessments (which were selected by examining rules for "poetry explication"; see, e.g.,[20]), participants also answered whether the poem rhymed, with choices "no, not at all," "yes, but badly," and "yes, it rhymes well."

As specified in our pre-registration (https://osf.io/82h3m), we predicted (1) that participants' assessments would be more positive when told the poem is human-written than when told the poem is AI-generated, and (2) that a poem's actual authorship (human or AI) would make no difference in participants' assessments. We also predicted that expertise in poetry (as measured by the self-reported experience with poetry) would make no difference in assessments.

Ratings of overall quality of the poems are lower when participants are told the poem is generated by AI than when told the poem is written by a human poet (two-sided Welch's $t(4571.552) = -17.398$, $p < 0.0001$, $p_{Bonf} < 0.0001$, $Mean_{difference} = -0.814$, Cohen's $d = -0.508$, 99.5% CI −0.945 to −0.683), confirming earlier findings that participants are biased against AI authorship[2,7,15]. However, contrary to earlier work[14,16,17] we find that ratings of overall quality are *higher* for AI-generated poems than they are for human-written poems (two-sided Welch's $t(6618.345) = 27.991$, $p < 0.0001$, $p_{Bonf} < 0.0001$, $Mean_{difference} = 1.045$, Cohen's $d = 0.671$, 99.5% CI 0.941 to 1.150); Fig. 1 compares the ratings distributions for AI-generated poems and human-written poems.

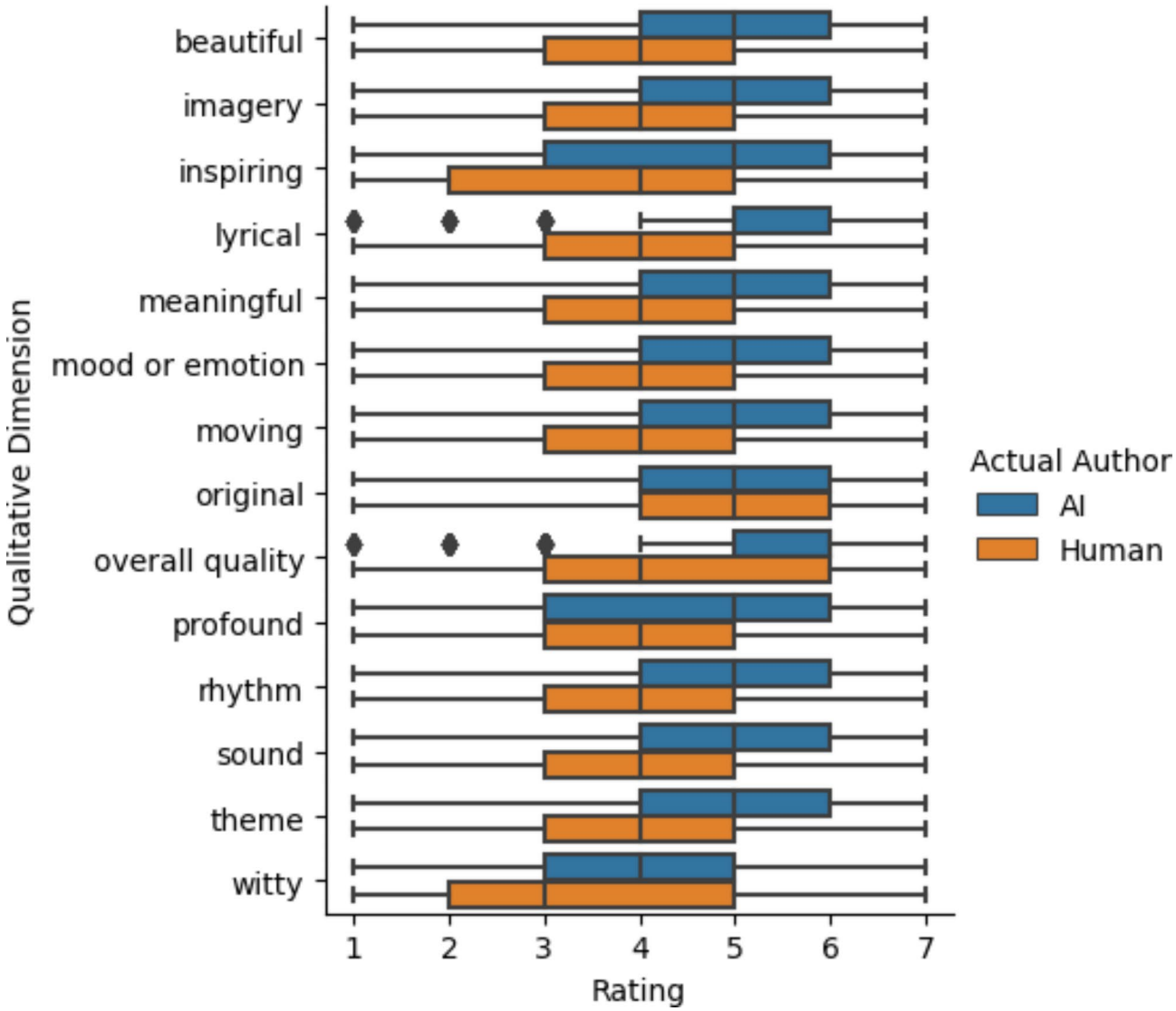

Fig. 1. Ratings for the 14 Measures of Poetic Excellence.

The same phenomenon – where ratings are significantly lower when *told* the poem is AI-generated but are significantly higher when the poem is *actually* AI-generated – holds for 13 of our 14 qualitative ratings. The exception is "original"; poems *are* rated as less original when participants are told the poem is generated by AI vs. being told the poem is written by a human (two-sided Welch's $t(4654.412) = -16.333$, $p < 0.0001$, $p_{Bonf} < 0.0001$, $Mean_{difference} = -0.699$, Cohen's $d = -0.478$, 99.5% CI −0.819 to −0.579), but originality ratings for actually AI-generated poems are not *significantly* higher than for actually human-written poems (two-sided Welch's $t(6957.818) = 1.654$, $p = 0.098$, $p_{Bonf} = 1.000$, $Mean_{difference} = 0.059$, Cohen's $d = 0.040$, 99.5% CI −0.041 to 0.160). The largest effect is on "rhythm": AI-generated poems are rated as having much better rhythm than the poems written by famous poets (two-sided Welch's $t(6694.647) = 35.319$, $p < 0.0001$, $p_{Bonf} < 0.0001$, $Mean_{difference} = 1.168$, Cohen's $d = 0.847$, 99.5% CI 1.075 to 1.260). This is remarkably consistent; as seen in Fig. 2, all 5 AI-generated poems are rated more highly in overall quality than all 5 human-authored poems.

We used a linear mixed effects model to predict the Likert scale ratings for each of our 14 qualitative dimensions. We used poem authorship (human or AI), framing condition (told human, told AI, or told nothing), and their interaction as fixed effects. As specified in our preregistration, we initially planned to include four random effects: random intercepts per participant, random slope of poem authorship per participant, random intercept per poem, and random slope of framing condition per poem. As in Study 1, we followed[19] in checking the models for overparameterization; PCA dimensionality reduction revealed that the models were overparameterized, specifically because of the random slopes for framing condition per poem. An attempt to fit a zero-correlation-parameter model did not prevent overparameterization; we therefore fit a reduced model for each DV without the random slopes for framing condition. ANOVA comparisons between the full and reduced models for each DV found that the reduced model provided at least as good a fit for 12 of the 14 DVs: all except "original" and "witty". We therefore proceed with the reduced model.

For 9 of our 14 qualities, human authorship had a significant negative effect ($p < 0.005$), with poems written by human poets rated lower than poems generated by AI; for 4 qualities the effect was negative, but merely suggestive ($0.05 < p < 0.005$). The only quality for which there is not even a suggestive negative authorship effect is "original" (b = -0.16087, SE = 0.10183, df = 29.01975, t = -1.580, $p = 0.1250$). For 12 of our 14 qualities, the "told human" framing condition had a significant positive effect, and poems are rated more highly when participants are told that the poem is written by a human poet; for "inspiring" (b = 0.21902, SE = 0.11061, df = 693.00000, t = 1.980, $p = 0.04808$) and "witty" (b = 0.28140, SE = 0.12329, df = 693.00024, t = 2.282, $p = 0.02277$) the effect is merely suggestive. For all 14 models, the explanatory power is substantial (conditional R-squared > 0.47). Detailed analysis for all qualities can be found in our supplementary materials.

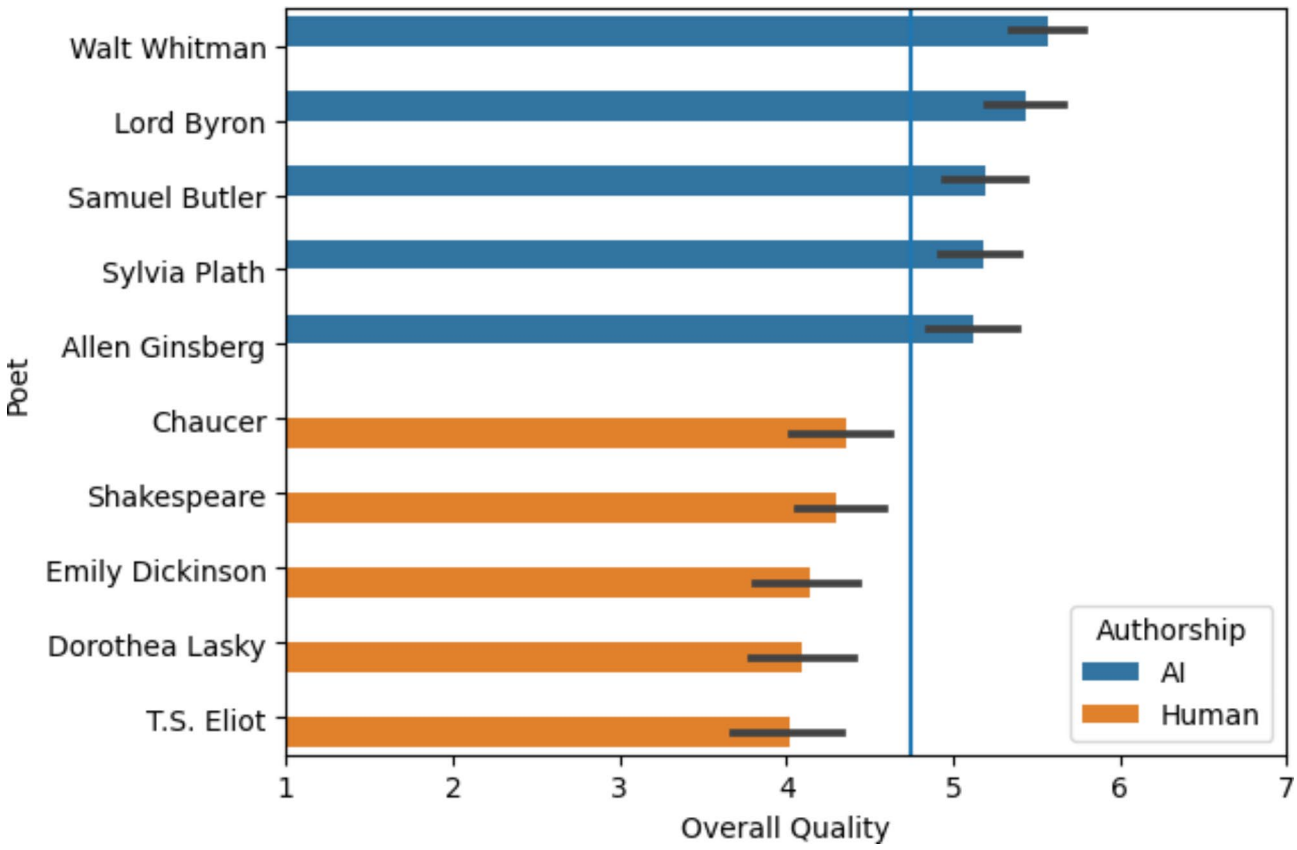

Fig. 2. Overall Quality Ratings for Study 2 Poems. (Error bars correspond to 99.5% confidence intervals. The vertical blue line corresponds to the mean rating across all poems and participants (4.7).)

Factor analysis of qualitative ratings

As specified in our pre-registration, we planned to factor analyze responses to the following scales: moving, profound, witty, lyrical, inspiring, beautiful, meaningful, original. However, we found higher-than-expected correlations among all of our qualitative ratings; polychoric correlations ranged from 0.472 to 0.886, with a mean of 0.77. Therefore, we performed factor analysis on all 14 qualitative ratings. Parallel analysis suggested 4 factors. We performed a maximum likelihood factor analysis with an oblique rotation; factor scores were estimated using the ten Berge method[21].

Factor 1 is most heavily weighted towards "beautiful," "inspiring," "meaningful," "moving," and "profound"; we take it to correspond to the poem's emotional quality, and call it "Emotional Quality." Factor 2 is most heavily weighted towards "rhythm," "lyrical," and "sound"; we take it to be the poem's formal, including structural or metrical, quality, and call it "Formal Quality." Factor 3 is most heavily weighted towards "imagery," "mood or emotion," and "theme"; we take it to reflect the poem's ability to capture a particular poetic "Atmosphere," and we call it "Atmosphere." Factor 4 is most heavily weighted toward "witty" and "original"; we take it to reflect how *creative* or *unique* the poem is, and we call it "Creativity." Fig. 3 shows the factor loadings for each qualitative dimension.

For each of the four factors, we used a linear mixed effects regression to predict factor values for each participants' rating of each poem, using the same fixed and random effects used for the 14 qualitative dimension DVs. We again found that the preregistered random effects overparameterized the models, and used the reduced models with no random slopes for framing condition.

We find that across all four factors, the explanatory power of the models is substantial (conditional R-squared > 0.5). The "told human" framing condition has a significant positive effect on all factors, and human authorship has a significant negative effect on 3 of the 4 factors. Figure 4 shows factor scores for human and AI authorship; Fig. 5 shows factor scores for each framing condition; the results for each of the 4 factor-prediction models, with the results for overall quality for comparison, can be found in Table 1.

Using qualitative ratings to predict discrimination

As in Study 1, we also used a mixed effects logistic regression (fit to a binomial distribution) to predict participant responses to the discrimination question ("written by a human" or "generated by AI") for participants in the "told nothing" framing condition. We included authorship (human or AI), stimulus line count (scaled), stimulus all-lines-rhyme, and stimulus first-person as fixed effects, with random intercepts for participants (dropping stimulus quatrain and stimulus first-person from the model we used in Study 1 due to high multicollinearity in Study 2-poem's smaller set of 10 poems). As expected, explanatory power of the model was low (conditional R-squared: 0.071, marginal R-squared: 0.013), but as in Study 1, we found that stimulus authorship (b = -0.435689, SE = 0.125832, z = -3.462, $p = 0.000535$) was once again significantly predictive of participants' responses: being written by a human poet *decreased* the likelihood that a participant would respond that the poem was written by

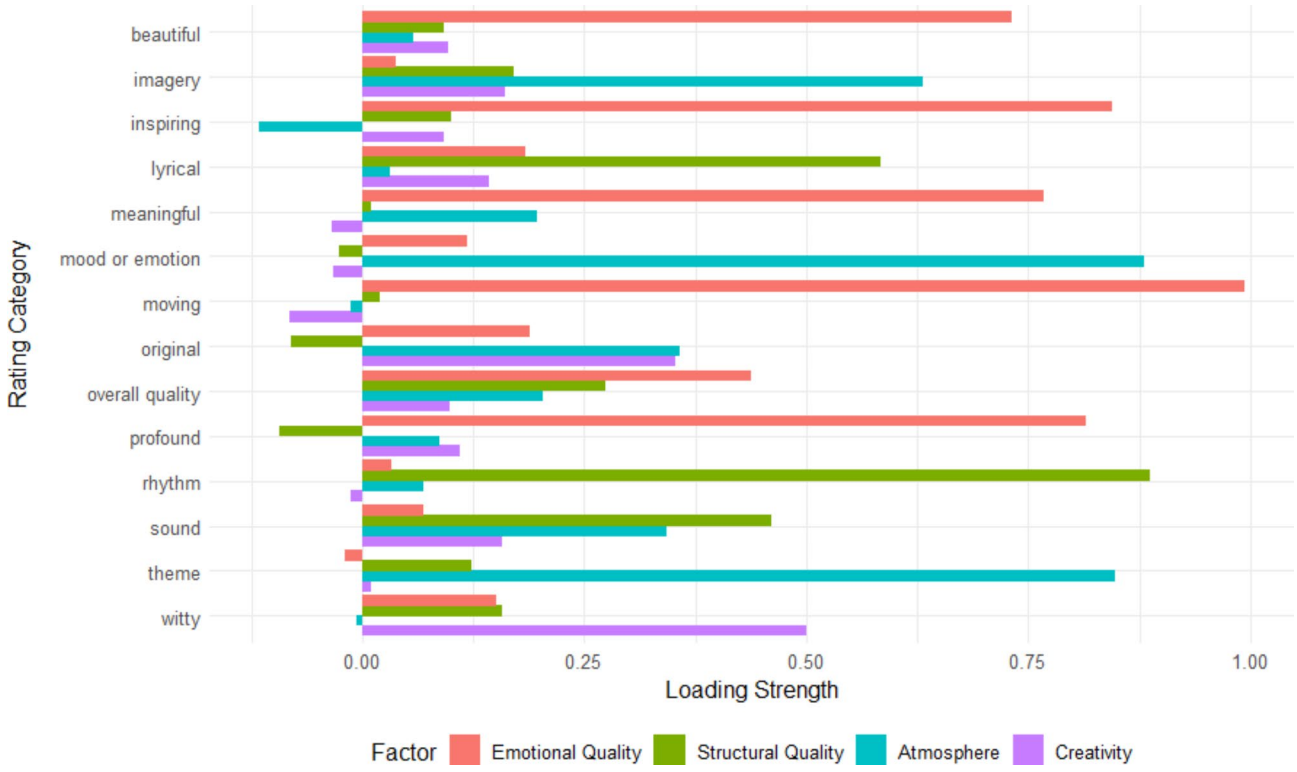

Fig. 3. Factor Loadings for each Qualitative Dimension.

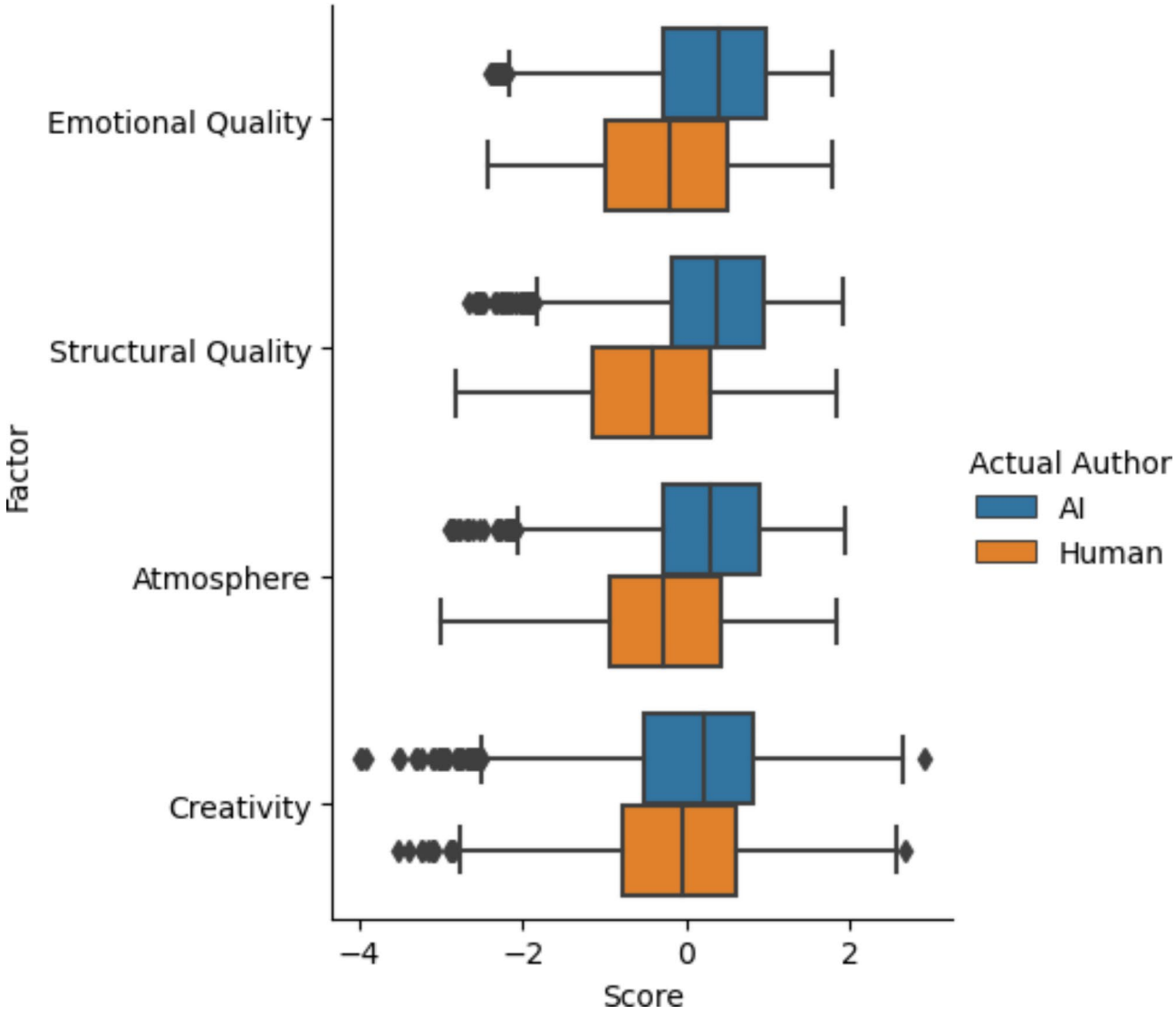

Fig. 4. Scores for the Four Factors for AI-Generated and Human-Written Poems.

a human poet, with the odds of a human-authored poem being judged human-authored less than two-thirds that of an AI-generated poem (OR = 0.647). This finding replicates the main result of our first study.

As an exploratory analysis, we also fit a model with our four factors Emotional Quality, Formal Quality, Atmosphere, and Creativity. We included authorship and these four factors as fixed effects, with random intercepts for participants. Effectively, this model replaces the structural features of the previous model (stimulus line count, stimulus all-lines-rhyme, and stimulus first-person) with qualitative features. The explanatory power of this model was higher (conditional R-squared: 0.240, marginal R-squared: 0.148), suggesting that qualitative features may have more influence than structural features on participants' beliefs about a poem's authorship. Atmosphere (b = 0.55978, SE = 0.11417, z = 4.903, $p < 0.0001$) was significantly predictive: higher scores for Atmosphere increased the likelihood that a participant predicted the poem was written by a human. We also found suggestive positive effects for Emotional Quality (b = 0.22748, SE = 0.11402, z = 1.995, $p = 0.04604$) and Creativity (b = 0.18650, SE = 0.07322, z = 2.547, $p = 0.01087$), suggesting that higher scores for Emotional Quality and Creativity may also increase the likelihood that participants predict a poem was written by a human poet. Importantly, in this model, unlike previous discrimination models, authorship has no negative effect (b = 0.23742, SE = 0.14147, z = 1.678, $p = 0.09332$). This suggests that the "more human than human" phenomenon identified in Study 1 might be caused by participants' more positive impressions of AI-generated poems compared to poems authored by human poets; when accounting for these qualitative judgments, the "more human than human" phenomenon disappears.

In summary, Study 2 finds that participants consistently rate AI-generated poetry more highly than the poetry of well-known human poets across a variety of factors. Regardless of a poem's actual authorship, participants consistently rate poems more highly when told that a poem is written by a human poet, as compared to being told that a poem was generated by AI. The preference for AI-generated poetry at least partially explains

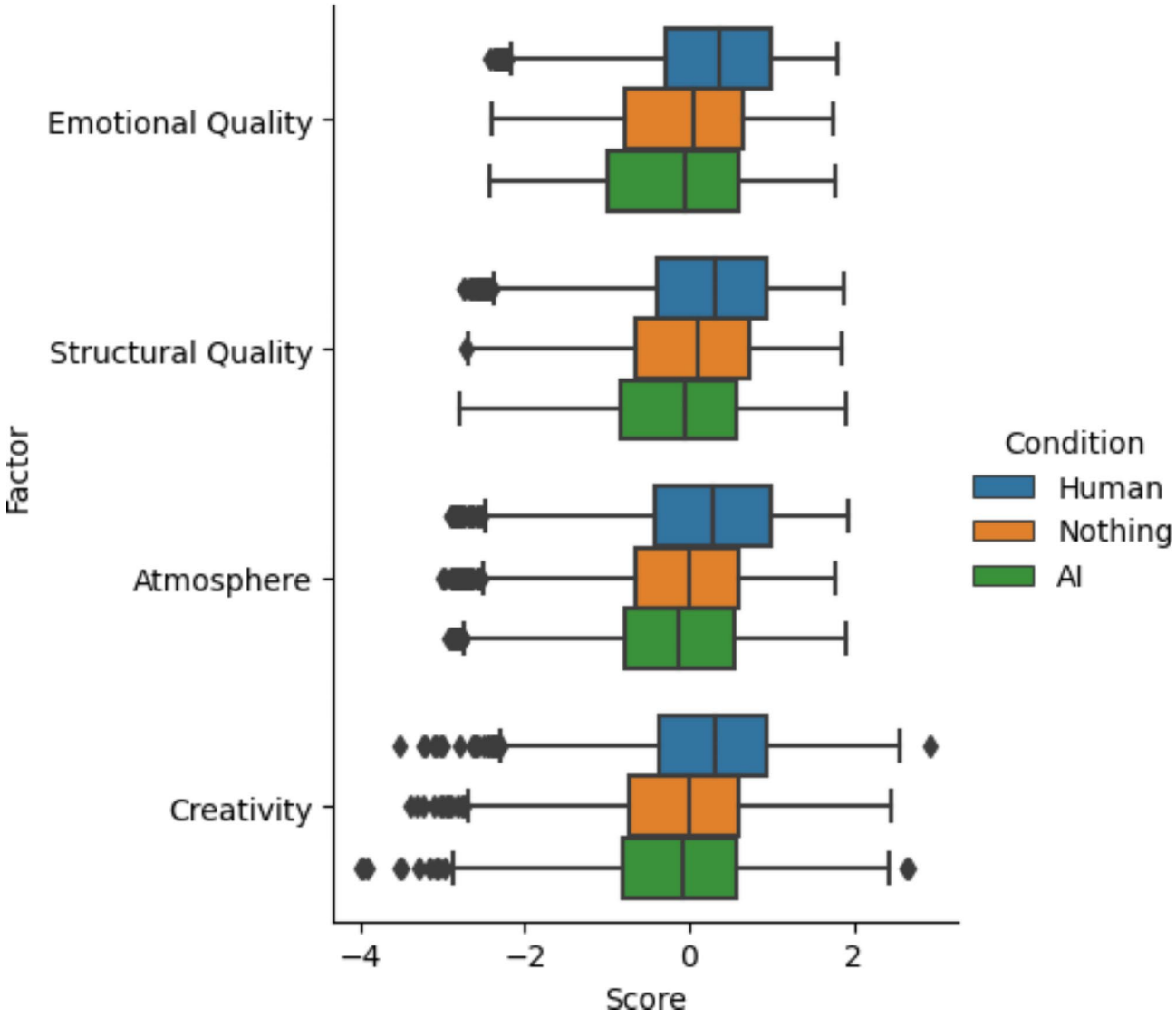

Fig. 5. Factor Scores for each Framing Condition.

the "more human than human" phenomenon found in Study 1: when controlling for participants' ratings, AI-generated poems are no longer more likely to be judged human.

Discussion

Contrary to what earlier studies reported, people now appear unable to reliably distinguish human-out-of-the-loop AI-generated poetry from human-authored poetry written by well-known poets. In fact, the "more human than human" phenomenon discovered in other domains of generative AI[1–5,10,11] is also present in the domain of poetry: non-expert participants are more likely to judge an AI-generated poem to be human-authored than a poem that actually is human-authored. These findings signal a leap forward in the power of generative AI: poetry had previously been one of the few domains in which generative AI models had not reached the level of indistinguishability in human-out-of-the-loop paradigms.

Furthermore, people prefer AI-generated poetry to human-authored poetry, consistently rating AI-generated poems more highly than the poems of well-known poets across a variety of qualitative factors. This preference at least partially explains the "more human than human" phenomenon: when controlling for people's opinions about the excellence of various aspects of poems such as their rhythmic quality, authorship no longer has a significant negative effect on beliefs about authorship, suggesting that people are more likely to believe that AI-generated poems are human-written *because* they prefer the AI poems and because they assume that they are more likely to like human-written than AI-generated poems.

So why do people prefer AI-generated poems? We propose that people rate AI poems more highly across all metrics in part because they find AI poems more straightforward. AI-generated poems in our study are generally more accessible than the human-authored poems in our study. In our discrimination study, participants use variations of the phrase "doesn't make sense" for human-authored poems more often than they do for AI-

	Dependent variable				
	Perceived as human-written				
	Emotional Quality	Formal Quality	Atmosphere	Creativity	Overall Quality
	(1)	(2)	(3)	(4)	(5)
Human Authorship	-0.561*** (-0.935, -0.187)	-0.835*** (-1.114, -0.555)	-0.614*** (-0.951, -0.277)	-0.222 (-0.588, 0.144)	-1.117*** (-1.518, -0.715)
Human Framing	0.242*** (0.053, 0.431)	0.256*** (0.103, 0.410)	0.291*** (0.122, 0.459)	0.230*** (0.025, 0.434)	0.409*** (0.141, 0.677)
AI Framing	-0.100 (-0.287, 0.087)	-0.070 (-0.222, 0.083)	-0.055 (-0.222, 0.112)	-0.078 (-0.281, 0.125)	-0.135 (-0.401, 0.130)
Authorship x Human Framing	0.174** (-0.011, 0.360)	0.148* (-0.049, 0.345)	0.156* (-0.036, 0.349)	0.160* (-0.028, 0.349)	0.380*** (0.043, 0.717)
Authorship x AI Framing	-0.059 (-0.243, 0.125)	-0.035 (-0.231, 0.160)	-0.076 (-0.267, 0.115)	-0.024 (-0.211, 0.163)	-0.159 (-0.493, 0.176)
Constant	0.216* (-0.065, 0.498)	0.338*** (0.135, 0.541)	0.217* (-0.031, 0.465)	0.040 (-0.242, 0.321)	5.300*** (5.003, 5.597)
Conditional R^2	0.590	0.542	0.537	0.506	0.536
Marginal R^2	0.107	0.192	0.128	0.038	0.149
Observations	6,960	6,960	6,960	6,960	6,960
Log Likelihood	-8,014.583	-8,211.183	-8,304.574	-8,532.946	-11,762.000
Akaike Inf. Crit	16,051.170	16,444.370	16,631.150	17,087.890	23,545.990
Bayesian Inf. Crit	16,126.490	16,519.690	16,706.480	17,163.220	23,621.320

Table 1. Regression Coefficients with 99.5% Confidence Intervals for 4 Factors and Overall Quality Ratings linear Mixed Effects Regression Models. *Note:* *p < 0.05; **p < 0.01; ***p < 0.005.

generated poems when explaining their discrimination responses (144 explanations vs. 29 explanations). In each of the 5 AI-generated poems used in the assessment study (Study 2), the subject of the poem is fairly obvious: the Plath-style poem is about sadness; the Whitman-style poem is about the beauty of nature; the Lord Byron-style poem is about a woman who is beautiful and sad; etc. These poems rarely use complex metaphors. By contrast, the human-authored poems are less obvious; T.S. Eliot's "The Boston Evening Transcript" is a 1915 satire of a now-defunct newspaper that compares the paper's readers to fields of corn and references the 17th-century French moralist La Rochefoucauld.

Indeed, this complexity and opacity is part of the poems' appeal: the poems reward in-depth study and analysis, in a way that the AI-generated poetry may not. But because AI-generated poems do not have such complexity, they are better at unambiguously communicating an image, a mood, an emotion, or a theme to non-expert readers of poetry, who may not have the time or interest for the in-depth analysis demanded by the poetry of human poets. As a result, the more easily-understood AI-generated poems are on average preferred by these readers, when in fact it is one of the hallmarks of human poetry that it does *not* lend itself to such easy and unambiguous interpretation. One piece of evidence for this explanation of the more human than human phenomenon is the fact that *Atmosphere* – the factor that imagery, conveying a particular theme, and conveying a particular mood or emotion load on – has the strongest positive effect in the model that predicts beliefs about authorship based on qualitative factor scores and stimulus authorship. Thus, controlling for actual authorship and other qualitative ratings, increases in a poem's perceived capacity to communicate a theme, an emotion, or an image result in an increased probability of being perceived as a human-authored poem.

In short, it appears that the "more human than human" phenomenon in poetry is caused by a misinterpretation of readers' own preferences. Non-expert poetry readers expect to like human-authored poems more than they like AI-generated poems. But in fact, they find the AI-generated poems easier to interpret; they can more easily understand images, themes, and emotions in the AI-generated poetry than they can in the more complex poetry of human poets. They therefore prefer these poems, and misinterpret their own preference as evidence of human authorship. This is partly a result of real differences between AI-generated poems and human-written poems, but it is also partly a result of a mismatch between readers' expectations and reality. Our participants do not expect AI to be capable of producing poems that they like at least as much as they like human-written poetry; our results suggest that this expectation is mistaken.

As generative AI models become both more capable and more common, it is unclear whether ordinary people's expectations for generative AI will catch up to the reality of generative AI. Heuristics that may serve readers well for one generative model or one generation of generative model may not generalize to other models. People could reliably distinguish the poetry of GPT-2 from human-written poetry[16]; our results show they cannot distinguish the poetry of ChatGPT-3.5.

Given people's difficulties identifying machine-written texts, and their apparent trust that AI will not generate imitations of human experience, it may be worthwhile for governments to pursue regulations regarding transparency in the use of AI systems. The White House[22] and the European Union[23] have recently proposed regulations for disclosing the use of AI systems to generate texts and images. However, there is evidence that users often ignore such disclosures[24], so it is unclear to what extent such regulations can help. Identifying effective disclosure methods is a difficult but urgent question.

Methods

Experiment Design. In Study 1, 1,634 participants were randomly assigned to one of 10 poets, and presented with 10 poems in random order: 5 poems written by that poet, and 5 generated by AI "in the style of" that poet. For each poem, participants answered a forced-choice prompt asking whether they thought the poem was written by a human or generated by an AI program. Participants then rated their confidence in their answer on a scale from 0-100, and were prompted to explain their answer if they wanted to. Following the discrimination task, participants provided demographic information and indicated their familiarity and interest with poetry.

In Study 2, 696 participants were randomly assigned to one of three framing conditions: "told human", in which participants were told that all poems were written by the professional human poet, regardless of actual authorship; "told AI", in which participants were told that all poems were generated by AI, regardless of actual authorship; and "told nothing", in which participants were not told anything about the poem's authorship. Regardless of framing condition, all participants were presented with the same 10 poems: 5 AI-generated, and 5 written by human poets. We followed prior studies[3,5,17] in asking participants to rate each poem on a Likert scale. For each poem, participants rated each poem's overall quality on a 7-point Likert scale from "extremely bad" to "extremely good"; rated the extent to which the poem had each of 8 different qualities on a 7-point Likert scale from "strongly disagree" to "strongly agree"; rated 3 qualitative features on a 7-point Likert scale from "terrible" to "excellent"; rated the extent to which the poem conveyed a specific theme and the extent to which it conveyed a specific mood or emotion on a 7-point Likert scale from "terribly" to "extremely well". Participants were asked whether the poem rhymes, with the choices "no, not at all", "yes, but badly" and "yes, it rhymes well". Participants in the "told nothing" framing condition then answered a forced-choice prompt asking whether they thought the poem was written by a human or generated by an AI program. Following the assessment task (and discrimination task, where applicable) participants provided demographic information and indicated their familiarity and interest with poetry.

The University of Pittsburgh Institutional Review Board approved the study protocols; all experiments were performed in accordance with all relevant guidelines and regulations. Informed consent was obtained from all participants in both studies. We preregistered both studies (https://osf.io/5j4w9, https://osf.io/82h3m) prior to data collection.

Collecting and Generating Poems. We chose 10 English-language poets: Geoffrey Chaucer, William Shakespeare, Samuel Butler, Lord Byron, Walt Whitman, Emily Dickinson, T.S. Eliot, Allen Ginsberg, Sylvia Plath, and Dorothea Lasky. We aimed to cover a wide range of genres, styles, and time periods. We collected a total of 50 poems: 5 poems for each of our 10 poets. Poems were collected from mypoeticside.com, an online poetry database. Poems for each poet were sorted by popularity; we selected poems that were outside of the top 10 most popular poems for that poet, and what were of reasonable length (less than 30 lines). We then generated a total of 50 poems using ChatGPT 3.5. The model was given a simple prompt: "Write a short poem in the style of <poet> ". The first 5 poems generated by that prompt were chosen.

Choosing Qualitative Features. For our assessment study (Study 2), we chose 15 qualitative features for participants to rate among those identified by[20]: overall quality, imagery, rhythm, sound, beautiful, inspiring, lyrical, meaningful, moving, original, profound, witty, convey a particular theme, convey a particular mood or emotion, and rhyme. We chose only qualities that were unambiguously good, so that higher ratings on the Likert scale were easily interpreted as more positive. We chose qualities that we hoped would cover a wide range of qualitative experiences that participants could have of the poem: a poem's structural quality (rhythm, rhyme), its emotional content (moving, convey a particular mood or emotion), its creativity (original, witty), its aesthetic features (beautiful, lyrical), and the extent to which it communicates meaning (meaningful, profound, convey a particular theme).

Predicting Responses. In our discrimination study (Study 1), we predicted that participants would be unable to distinguish AI-generated poetry from human-written poetry. We based this on the fact that human-in-the-loop AI-generated poetry generated by GPT-2 had been shown to be indistinguishable from human-written poetry[16]; we predicted that poetry generated by ChatGPT 3.5 would be at least as good as the human-selected best poems generated by GPT-2. In our assessment study (Study 2), we predicted that participants' assessments would be more positive when told the poem was written by a human poet, compared to when participants are told the poem was generated by AI. We based this on similar findings in AI-generated art[2,7]. We also predicted that participant assessments would not significantly differ between AI-generated poetry and human-written poetry, based on the fact that participants in our discrimination study had not been able to reliably distinguish AI-generated poetry from human-written poems. In both studies, we predicted that expertise in poetry would not make a difference.

Participant Recruitment. For Study 1, we recruited a sample of 1,634 US-based participants through Prolific. Participants had a median age of 37; 49.6% were male, 48.5% female, and 1.9% non-binary or prefer not to say. They were paid $1.75 ($13.07/hr). For Study 2, we recruited 696 US-based participants through Prolific. Participants had a median age of 40; 50.4% were male, 46.6% female, and 3% non-binary or prefer not to say. They were paid $2.00 ($11.99/hr).

Limitations. Our results are limited to the most recent generation of generative language models, and to people's current beliefs and biases regarding AI-generated texts. It is likely that as new generative language models are created and as AI-generated texts become more prevalent, what "sounds human" in a poem or other piece of text will change. In particular, it is possible that expectations regarding the qualitative differences between AI-generated text and human-authored text may change over time.

Significance statement

We show that, in contrast to previous studies, people are now unable to distinguish AI-generated poetry from the poetry of well-known human poets, being more likely to judge AI-generated poems to be human-written

and rating AI-generated poetry more highly along several aesthetic dimensions. We explain this by appealing to people's mistaken expectations about what AI are able to do and to their own aesthetic preferences. Poetry was previously one of the last remaining domains of text in which generative AI language models had not yet reached this level of indistinguishability; our findings indicate that the capabilities of generative AI models have outpaced people's expectations of AI, even as the use of generative language models like ChatGPT have become increasingly commonplace.

Data availability

All data and code used in analysis is available at an OSF repository: https://osf.io/by4cg/files/osfstorage.

Received: 13 July 2024; Accepted: 17 October 2024
Published online: 14 November 2024

References

1. Sun, Y., Yang, C.-H., Lyu, Y. & Lin, R. From pigments to pixels: A comparison of human and AI painting. *Appl. Sci.* **12**, 3724. https://doi.org/10.3390/app12083724 (2022).
2. Ragot, M., Martin, N. & Cojean, S. AI-generated vs. human artworks. A perception bias towards artificial intelligence? In *CHI Conference on Human Factors in Computing Systems*, 1–10 (2020). https://doi.org/10.1145/3334480.3382892.
3. Nightingale, S. J. & Farid, H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl. Acad. Sci.* **119**, 1–3. https://doi.org/10.1073/pnas.2120481119 (2022).
4. Tucciarelli, R., Vehar, S. & Tsakiris, M. On the realness of people who do not exist: The social processing of artificial faces. *iScience* **25**, 105441 (2022).
5. Miller, E. J. et al. AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychol. Sci.* **34**(12), 1390–1403. https://doi.org/10.1177/09567976231207095 (2023).
6. Gorenz, D. & Schwarz, N. How funny is ChatGPT? A comparison of human- and A.I.-produced jokes (2024). https://doi.org/10.31234/osf.io/5yz8n.
7. Bellaiche, L. et al. Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cogn. Res. Princ. Implic.* **8**, 42. https://doi.org/10.1186/s41235-023-00499-6 (2023).
8. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
9. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* (2023). https://doi.org/10.48550/arXiv.2307.09288.
10. Jakesch, M., Hancock, J. T. & Naaman, M. Human heuristics for AI-generated language are flawed. *Proc. Natl. Acad. Sci.* **120**, e2208839120 (2023).
11. Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S. & Smith, N. A. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) 7282–7296 (Association for Computational Linguistics, Cedarville, OH, 2021) (2021).
12. Elam, M. Poetry will not optimize: or, What is literature to AI?. *Am. Lit.* **95**, 281–303 (2023).
13. Linardaki, C. Poetry at the first steps of Artificial Intelligence. *Humanist Stud. Digit. Age* https://doi.org/10.5399/uo/hsda/7.1.6 (2022).
14. Gunser, V., Gottschling, S., Brucker, B., Richter, S., Çakir, D. & Gerjets, P. The pure poet: How good is the subjective credibility and stylistic quality of literary short texts written with an artificial intelligence tool as compared to texts written by human authors?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44) (2022). https://escholarship.org/uc/item/1wx3983m.
15. Rahmeh, H. Digital verses versus inked poetry: Exploring readers' response to AI-generated and human-authored Sonnets. *Sch. Int. J. Linguist. Lit.* **6**(9), 372–382 (2023).
16. Köbis, N. & Mossink, L. D. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput. Hum. Behav.* **114**, 106533. https://doi.org/10.1016/j.chb.2020.106553 (2021).
17. Hitsuwari, J., Ueda, Y., Yun, W. & Nomura, M. Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Comput. Hum. Behav.* **139**, 107–502. https://doi.org/10.1016/j.chb.2022.107502 (2023).
18. Benjamin, D. J. et al. Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10. https://doi.org/10.1038/s41562-017-0189-z (2018).
19. Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. Parsimonious mixed models. *arXiv* (2015). https://doi.org/10.48550/arXiv.1506.04967.
20. Kennedy, X. J. & Gioia, D. *An Introduction to Poetry* 13th edn. (Pearson, 2009).
21. ten Berge, J. M. F., Krijnen, W. P., Wansbeek, T. & Shapiro, A. Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra Appl.* **289**, 311–318 (1999).
22. Nelson, A., Friedler, S. & Fields-Meyer, F. Blueprint for an AI bill of rights: A vision for protecting our civil rights in the algorithmic age. *White House Office of Science and Technology Policy* (2022). (18 October 2022).
23. European Commission. Proposal for a regulation laying down harmonised rules on artificial intelligence. *Shaping Europe's Digital Future* (2021). (18 October 2022).
24. Acquisti, A., Brandimarte, L. & Hancock, J. How privacy's past may shape its future. *Science* **375**, 270–272 (2022).

Author contributions

B.P. and E.M. wrote the main manuscript text. All authors reviewed the manuscript.

Declarations

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/s41598-024-76900-1.

Correspondence and requests for materials should be addressed to B.P.

Reprints and permissions information is available at www.nature.com/reprints.