Journal of High School Science

Self-reported mentalizing and AI differentiation: An empirical study with implications for construct validity

Sankarlingam S

Submitted: June 1, 2026, Revised: version 1, July 1, 2026, version 2, July 2, 2026, version 3, July 2, 2026
Accepted: July 4, 2026

Abstract
Generative artificial intelligence (gAI) systems are increasingly capable of producing human-like writing, yet little is known about the cognitive factors that influence an individual's ability to distinguish AI-generated from human-authored text. This study investigated whether self-reported mentalizing, measured using the Multidimensional Mentalizing Questionnaire (MMQ), was associated with performance on an AI differentiation task involving scientific abstracts and news headlines generated using a contemporary large language model. To evaluate this relationship, multiple complementary analytical approaches were employed, including Pearson correlation, multiple linear regression, principal component analysis, bootstrap resampling, cross-validation, and model comparison. Participants performed near chance overall, and no MMQ composite, individual factor, or latent component demonstrated a robust association with AI differentiation performance after correction for multiple comparisons and evaluation of out-of-sample generalizability. Although one MMQ factor (Reflexivity) showed a nominal association with AI news headline differentiation, this relationship was not robust across validation procedures.These findings indicate that self-reported mentalizing, as operationalized by the MMQ, did not reliably predict performance on the present AI differentiation task. However, the results should be interpreted cautiously because AI differentiation itself remains an emerging construct. Unlike many psychological tasks that involve relatively stable human behaviors, AI differentiation depends on identifying outputs from rapidly evolving algorithmic systems whose stylistic characteristics change across model generations. Consequently, it remains uncertain whether AI differentiation represents a stable psychological construct with enduring cognitive determinants or performance against a temporally specific technological artifact. The present findings therefore apply to the specific experimental paradigm and LLM generation examined and highlight the need for future research to establish the construct validity and temporal stability of AI differentiation before drawing broader conclusions regarding its cognitive predictors.

Keywords
Large Language Models, Generative Artificial Intelligence, Human–AI interaction, Theory of Mind, Mentalizing, AI detection, Individual differences, Psychological constructs, Construct validity, Cognitive psychology

Shivaani Sankarlingam, Enloe High School, Wake County Public School System, 128 Clarendon Crescent, Raleigh NC 27610, USA. shivaani.sankarlingam@gmail.

https://doi.org/10.64336/001c.164666

1. Introduction

Since the publication of Turing's famous "Imitation Game" paper in the 1950s, much work has been done in the realm of AI and its ability to pass as authentically 'human' (1). AI has shown promise in fields such as healthcare and finance (2), but simultaneously raises many concerns, such as its ability to hallucinate (parrot untrue ideas) and spread misinformation (3, 4); its ability to pass as student work and propagate plagiarism (5-7). In the scientific domain, particular concern has focused on the ability of AI-assisted writing to generate scientific abstracts that are difficult for peer reviewers to distinguish from those written by humans (7-10). To further these concerns, AI is only increasing in sophistication: Large Language Models (LLMs) such as ChatGPT are capable of unexpected tasks that are not explicitly coded into them, displaying ostensibly human intellectual abilities (11).

Currently, there is a substantial corpus of work dissecting the human ability to distinguish between AI and human writing, but very little of it is completely directed towards understanding the cognitive factors that affect an individual's ability to differentiate between the two (12). Given the rapid advancement and increasing integration of generative AI into scientific communication, it is essential to better understand the cognitive mechanisms that may underlie humans' ability to distinguish AI-generated from human-authored writing, while recognizing that the stability and validity of this ability as a psychological construct remain to be established.

Thus, in the present study, prior investigation into this field will be continued while guided by the overarching question: *Is there a relationship between an English-proficient adult's perceived mentalizing skill and their ability to differentiate between human and AI writing?* Beyond this, this paper will explore the further, more exploratory idea: *Which representation of self-reported mentalizing could best explain individual differences in AI differentiation ability?*

1.1 Differentiation between human and generative AI (gAI) writing

1.1.1 *Previous studies*

Most existing studies have focused on whether humans can accurately distinguish AI-generated from human-authored text. This question has been tested on a variety of AI-generated writing, including—but not limited to—news headlines (12), scientific abstracts (13), poetry (14), student essays (5), recipes (15), and more. The results in these studies form a consensus that human abilities to discern AI outputs are rarely above chance (50%), with variations caused by evaluators' AI expertise and familiarity (10, 15, 16).

Additionally, work has also been done to understand the abilities of AI itself to correctly attribute text origin. Studies by Ippolito and colleagues have shown that both humans and AI systems make errors when distinguishing AI-generated from human-authored text; however, the underlying causes of these errors differ and tend to emerge under different circumstances (17, 18). Overall, AI-detection machines are consistently high-scoring, whereas human scores fluctuate frequently, with some evaluators scoring as high as 85% and others scoring less than chance (18, 19), emphasizing

the hypothesis that there are underlying factors leading to such varied results among humans.

1.1.2 *Flawed strategies for AI discrimination*

Research into the patterns that human evaluators pursue to differentiate between human and AI excerpts reveals that humans rely on a multitude of heuristics, or mental strategies, to attempt this task, most of which are somehow flawed or incomplete (16, 20). In a study determined to understand the capabilities of non-experts to differentiate between humans and AI in a variety of textual media, evaluators focused on the composition of a text (grammar, tone, clarity, or "template"-esque writing), rather than the content of the text itself, to make judgements about its origin (15). The behavior of participants in Clark's study contrasts Hadan et al.'s findings about the actual differences between gAI and human outputs, which show that AI-written papers lack the insight and research details found in human-written papers, a nuance that is difficult to discern for most non-experts (10).

One of the first studies to systematically study the flawed heuristics used by humans to complete differentiation tasks was a landmark 2023 study by Jakesch et al., which analyzed the human judgments and thought processes behind attributing AI origin to a text conversation. Their results revealed that human evaluators would often rely on features like first-person pronouns, contractions, or family topics to attribute human origin, reaffirming the human overreliance on structure over content found in the studies mentioned previously (16). Despite relatively low scores in differentiation tasks such as these, evaluators in most studies seem to display consistently high levels of confidence in their perceived accuracy, regardless of their actual performance (5, 15, 20). This trend also indicates that a diminished ability to differentiate is not necessarily due to uncertainty, but rather some other underlying factor that individuals are not consciously aware of when encountering AI in unexpected contexts.

1.1.3 *Individual differences in AI differentiation ability*

Though there is a paucity of literature surrounding the cognitive and psychological processes that lend themselves to the ability to differentiate between human and AI works (21), some studies have found extraneous factors that may be able to predict or increase AI differentiation ability. Research has already found that gender (22), overconfidence in detection ability (17), and the use of incentives for accuracy (21) are all viable effectors of AI differentiation ability. Despite expertise in the relevant field and AI exposure also being traits commonly associated with higher attribution accuracy (10), certain works have contradicted this idea: in a study conducted with art and non-art majors, there were no statistically significant differences in AI differentiation ability for artworks (23), and there was only a marginal difference between English teachers and high school students in another later study (6). These findings suggest a strong probability that there is an interplay of multiple factors, beyond simply expertise, dictating AI discrimination ability. Among current research, there is a small sphere of works addressing the psychological factors leading to greater discrimination accuracy. One of these studies, which focused on AI-generated poetry, found nearly no correlation with empathy but did find a strong

correlation between animism (the inclination to anthropomorphize attributes), and differentiation skill (32). Another study, which directly elaborated on the findings of Histuwari's work, focused on cognitive and affective empathy (which neared statistical significance in the previous study), executive functioning, and non-verbal fluid intelligence (the ability to solve novel visual problems without prior information) in the context of news headlines and social media comments. Its results showed strong promise in overall executive functioning in the social media comment task and, especially, non-verbal fluid intelligence for all AI media (12). These studies strengthen the interest in investigating cognitive abilities as a predictor of AI differentiation ability.

1.2 The current gap in the literature

Chein's findings in the aforementioned study have been used as foundational knowledge for the current one, which aimed to extend their findings. As stated in the paper, no studies before had explored the central cognitive skills that contributed to AI differentiation ability, with their study partially bridging this gap by investigating empathy, non-verbal fluid intelligence, and executive functioning. However, many cognitive skills were left unaddressed, with the paper itself stating that "other…aspects of executive control…and mentalizing skill" may also affect AI Discrimination (12). The current study takes into account this gap and aims to bridge it by investigating the effects of mentalizing on differentiation ability.

1.3 Mentalizing skill

Mentalizing, the cognitive factor of interest in this study, is defined as the ability to represent others and oneself in terms of internal mental states (24). It also correlates positively with fluid intelligence and is linked strongly to empathy, two skills that have been discussed at length in similar studies (12). By focusing on mentalizing, inquiry will be extended into regions that have shown promise in previous studies.

Gori et al. proposed that mentalizing as a whole can be represented by four 'polarities', or aspects—explicit-implicit, outside-inside, self-other, and cognitive-affective—which all detail different components of mentalizing such as demands on mental effort, internal vs. external processes and behaviors, empathy, and rationality (25). This hypothesis is backed primarily by Fonagy's four-dimensional model (26) and is what led to the further distinction between good and poor mentalizing by Vera Cruz et al. According to them, good mentalizing demands that an individual be able to use the four polarities dynamically, balancing them appropriately by circumstance, whereas poor mentalizing is defined by a distinct lack of this mental flexibility (27). These findings suggest that mentalizing should be treated not as a single cognitive skill, but as the interplay of many distinct factors, each with different implications and effects on an individual's cognitive ability. Therefore, this study will treat both good and poor mentalizing, as well as the many factors of mentalizing ability, as separate predictors of interest.

1.4 Study objectives and hypotheses

This study aimed to further understand the cognitive skills affecting AI differentiation,

hypothesizing that group-level ability will be around chance, with high variation in individual scores. The study examined multiple dimensions of self-reported mentalizing in the context of AI-generated scientific abstracts and news headlines, testing the central hypothesis (H_a) that the adaptive dimensions of mentalizing (good mentalizing), which have previously been associated with fluid intelligence, would also be positively associated with AI differentiation ability.

However, despite prior work suggesting that mentalizing may contribute to distinguishing human and AI text, it remains unclear which aspects of mentalizing are most relevant. Broad composite measures, such as the MMQ, may obscure relationships among individual dimensions, whereas alternative representations —such as latent components or multivariate models—may better capture the involved cognitive processes. Accordingly, this study compared several theoretically and statistically motivated representations of self-reported mentalizing to determine which best explains the differences in AI differentiation performance.

2. Methods

2.1 Participants

Participants were recruited via an online recruiting platform (28, 29) and were required to be legal adults (18 years or older), English-fluent or native, and have no formal training or education in AI. These requirements were all chosen to limit the confounding factors related to AI differentiation and ensure that scores would be based purely on a participant's inherent ability to distinguish between human and AI English writing. In total, 110 participants were recruited, but 3 were excluded for failing to meet these requirements. The final sample consisted of 107 individuals (M age = 29.1 years, SD = 10.2 years; 37 Male, 66 Female, 4 Other). All participants were provided with and required to sign an electronic informed consent form prior to taking part in the study.

2.2 Stimuli

Two types of texts were used in the study: scientific research abstracts and news headlines. Excerpts were generated using ChatGPT's most recent model, at the time of writing, GPT-5.2 (30). Rather than using excerpts from previous studies, all texts were re-generated for the purpose of this study to accurately capture the current state and abilities of gAI. These texts were used to measure participants' AI discrimination ability in the differentiation task. All stimuli can be found in Appendix 1.

2.2.1 *Human and AI-generated texts*

Research abstracts were pursued as an excerpt type because of their medium length (longer generated texts are more prone to AI hallucinations), practicality (ChatGPT has a limited number of word tokens that can be used) (8), and use in other similar studies (7). Studies have shown that it is advisable to consider excerpts of various lengths, which is why abstracts were used to represent medium-length texts, and, later on, news headlines were used for short texts (17).

Five human-written research abstracts in 5 scientific disciplines (Psychology, Physics, Engineering, Biology, Medicine) were obtained from various peer-reviewed journals. These abstracts were then rephrased by ChatGPT (31),

which was given a prompt similar to that used by Gao et al.: "Please write three 200-300-word scientific abstracts for the article [title], in the [discipline] discipline, in the style of [journal] at [link]" (13). This was done to prevent AI hallucinations or fabrications. For each prompt, one of the 3 generated abstracts was chosen via a random number generator (1 to 3) for use in the differentiation task. Randomization was pursued to increase the reliability of AI excerpt quality and characteristics.

As professional-level texts are amongst the most difficult to identify for non-professionals, news headlines were used so that expertise would not become a confounding factor (17). To prevent the possibility of the sourced news articles having been written or influenced by AI themselves, or falsification of news stories, all articles were extracted from a previously utilized, trustworthy news organization, the New York Times (12). To avoid the instance of the AI copying from its training data, efforts were made to use more recently published articles (15), but, simultaneously, to prevent participants from recognizing headlines, less covered stories from the previous year (2025), or ones that had lost their traction, were used (12). Similar to the scientific abstracts, 5 human-written news headlines were obtained and input to ChatGPT to be rephrased using the prompt "Please write 3 news headlines for the news article [title], in the style of the NYT at [link]" (13). One article of three was again chosen using a random number generator for the previously stated reasons.

2.3 Procedure

Participants were given a series of pre-validated and previously used behavioral tasks and self-report questionnaires. One behavioral task was used to assess AI discrimination ability (differentiation task), and one self-report questionnaire, the Multidimensional Mentalizing Questionnaire (MMQ), was used to quantify both good and poor mentalizing capacity (27). The order of administration was always the differentiation task first, followed by the MMQ.

The individual scores for the differentiation task (which measured AI differentiation ability for the specific conditions of the study) were compared to the scores for MMQ (for both good and poor mentalizing) using a Pearson correlational analysis. The AI differentiation task, which was comprised of two different types of texts, was also split such that each type of stimulus (scientific abstract and news headline) could be considered against MMQ scores. This made it possible to determine whether a certain factor/aspect of mentalizing was especially influential for one type of text. The correlation coefficients and graphs derived from this method were used to understand and quantify the strength of the relationship, if present, between each variable. The strength and direction of these correlations was used to answer the question of whether Mentalization was related in any way to AI discrimination ability. Additionally, some exploratory analyses were conducted to further evaluate whether alternative representations of mentalizing provided greater explanatory value for the question at hand.

2.3.1 *Differentiation task*

Participants were given an AI-differentiation task via Google Forms. The task had a total of 10 questions: 5 questions using abstracts and 5

using news headlines, which were collected from ChatGPT using the method discussed previously. The length of the task was intentionally kept brief to adhere to the procedures of previous literature and prevent respondent fatigue (6). Excerpts were presented side-by-side to limit the occurrence of guesswork; this format usually correlates with higher overall task scores (6). The order of AI-generated and human excerpts was determined by a digital coin toss, where tails indicated AI being presented first, and vice versa. This was also done to prevent informed guesswork. For each pair of excerpts, participants were asked to determine which was AI-generated and indicate their decision by pressing the number choice that matched their judgment (i.e., 1 for excerpt one, 2 for excerpt two).

2.3.2 *Multidimensional Mentalizing Questionnaire*

The Multidimensional Mentalizing Questionnaire (MMQ), developed to bypass the limitations of interview-based quantifiers, has been validated with good internal validity by its creators as a measure of perceived mentalizing ability in both Italian and English, though it assumes that participants can accurately report their own mentalizing tendencies (27). The MMQ requires participants to answer a series of 33 qualitative self-report questions ranking qualities of their emotional mental processes and interactions with others on a Likert scale of 1-5. The questionnaire is split into 6 factors (F1-F6), with scores in F1-3 (F1 = reflexivity; F2 = ego-strength; F3 = relational attunement) being directly correlated with good perceived mentalizing ability (high scores indicate high ability) and scores in F4-6 (F4 = relational discomfort; F5 = distrust; F6 = emotional dyscontrol) being inversely correlated (high scores indicate low ability). The full assessment can be found in Appendix 2. The structure of the MMQ is illustrated in Figure 1.

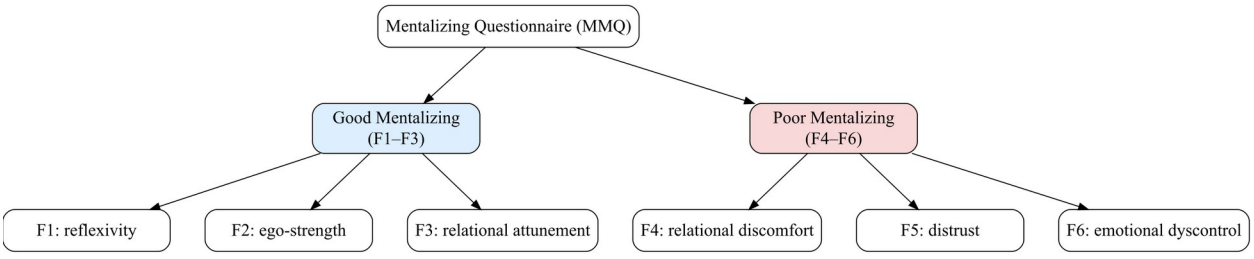

Figure 1. Visual model of Multidimensional Mentalizing Questionnaire factors.

This study utilized both F1-3 and F4-6, creating a separate linear regression for each factor set and AI differentiation. Doing so clearly illustrated the individual differences between good and poor perceived mentalizing ability as distinct predictors of AI attribution, and is supported by Vera Cruz's finding that the F1-3 were significantly correlated with each other and significantly negatively associated with F4-6, indicating accuracy in determining whether an individual believed that they possessed stronger or weaker mentalizing abilities (27). Gori et al. and Vera Cruz et al. both conceptualize F1-F3 and F4-F6 as higher-order

dimensions representing adaptive and maladaptive mentalizing (25, 27), and the present study therefore adopted this theoretically motivated grouping for its primary analyses while also examining each factor individually, in the event that one factor was disproportionately associated with AI differentiation ability.

2.3.3 *Exploratory Analyses*

However, because the prior analyses did not reveal any significant associations between the higher-order good and poor mentalizing scores and AI differentiation performance, several exploratory analyses were conducted to consider alternative explanations for the relationship between the various aspects of mentalizing and the dependent variable. These analyses included a multiple linear regression using all six MMQ factors simultaneously, a Principal Component Analysis (PCA) to identify latent dimensions of mentalizing, and bootstrap and cross-validation analyses to evaluate the stability and generalizability of any observed relationships.

Additionally, to evaluate competing explanatory frameworks, three nested linear models' explanatory power were directly compared using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). All models were correlated to AI News differentiation score (as it was the only type of AI differentiation to display a raw significant correlation with any aspect of mentalizing), but Model 1 included only good mentalizing, Model 2 added poor mentalizing, and Model 3 included all 6 MMQ factors. These analyses were all exploratory, and were intended to generate potential hypotheses for future research rather than confirm the primary hypothesis.

3. Results

3.1 AI differentiation Ability

First, overall AI differentiation ability was calculated as an extension of previous literature. The overall average accuracy of participants when differentiating between human and AI texts was 52.7%, while the averages for scientific abstracts and news headlines were 55.5% and 49.9%, respectively. When considering the results of the AI differentiation task at the $\alpha = 0.05$ significance level using a t-test, participants did not score significantly higher than chance overall ($t(106) = 1.53598$, $p = 0.06376$) and for the news headlines ($t(106) = -0.03245$, $p = 0.51293$), but did score significantly higher for the scientific abstracts ($t(106) = 2.05264$, $p = 0.02129$).

3.2 Effects of perceived mentalizing skill on differentiation

3.2.1 *Good mentalizing*

Individual scores for good mentalizing were calculated by finding the sum of the first 3 factors (F1-F3) of the MMQ ($M = 68.16$, $SD = 8.1$). Graphical representations of F1-3 scores compared to the various AI differentiation tasks are shown in Figure 2. All correlations were analyzed at the $\alpha = 0.05$ significance level using a t-test. Overall, good mentalizing was shown to have a statistically insignificant negative correlation to overall AI differentiation scores (Figure 2, A) ($r = -0.12$, $p = 0.2149$). The relationship between F1-F3 and the scores for AI scientific abstract differentiation (Figure 2, B) was also shown to be statistically

insignificant (r = -0.13, p = 0.1962). The correlation between F1-F3 and AI news headline differentiation (Figure 2, C) was found to be similar (r = -0.03, p = 0.742). These findings suggest that, overall, there is no significant correlation between AI differentiation ability and good mentalizing.

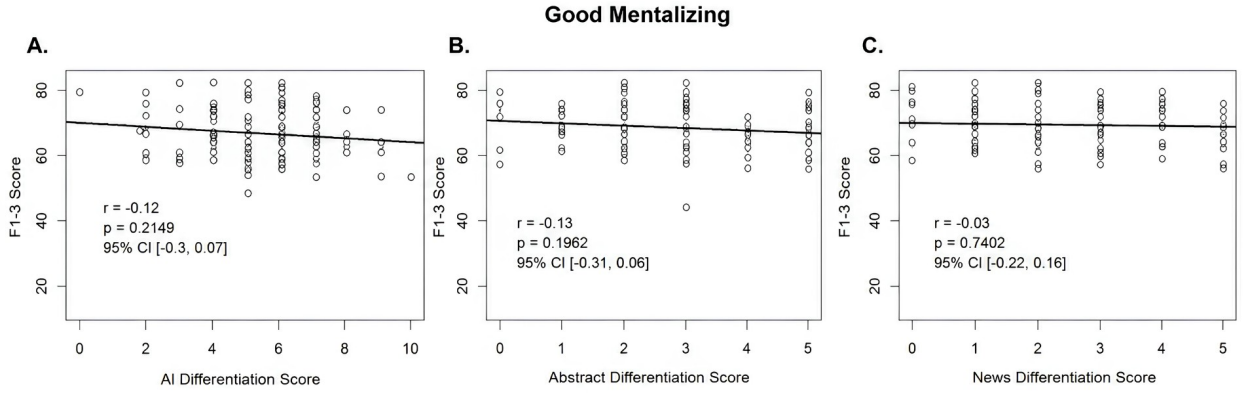

Figure 2. Relationship between F1-F3 (good mentalizing) scores and (**A**) AI differentiation (**B**) scientific abstract differentiation and (**C**) news headline differentiation.

3.2.2 *Poor mentalizing*

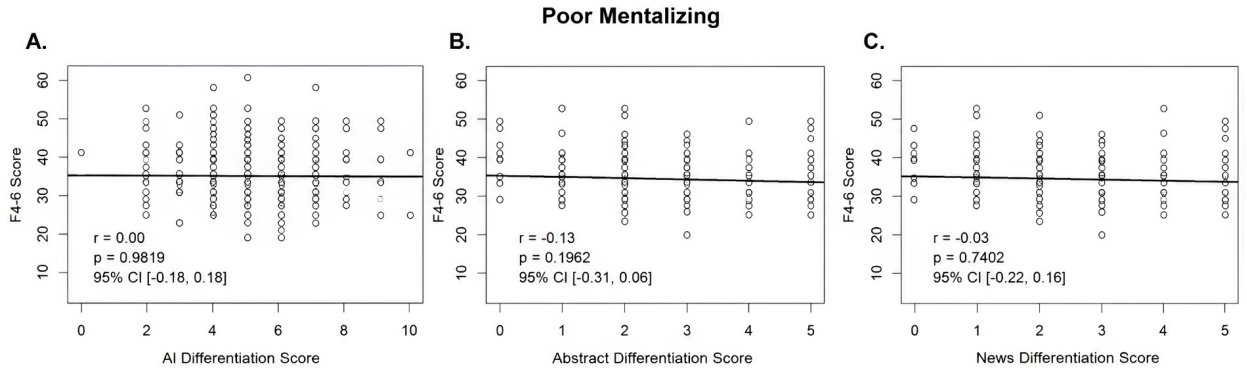

Figure 3. Relationship between F4-F5 (poor mentalizing) scores and (**A**) AI differentiation (**B**) scientific abstract differentiation and (**C**) news headline differentiation.

Individual scores for Poor mentalizing were calculated by finding the sum of the last 3 factors (F4-F6) of the MMQ (M = 36.4, SD = 7.72), and were shown to have no correlation to AI differentiation score (Figure 3, A) (r = 0.00, p = 0.9819). The relationships with AI news headlines (Figure 3, B) (r = -0.03, p = 0.7402) and AI abstract differentiation (Figure 3, C) (r =

-0.13, p = 0.1962) were also found to be statistically insignificant. These results suggest that the overall measure of poor mentalizing is unrelated to AI differentiation ability.

3.2.3 *Individual factor scores*

The relationships between each individual factor score and AI differentiation were also explored through separate correlation tests. Overall, most correlations found were statistically insignificant ($\alpha = 0.05$, $p > 0.05$). It was also found that r tended to be negative when relating factors that indicated good mentalizing, but positively correlated with those that indicated poor mentalizing. This pattern applies broadly to the findings, though the correlations with F5 were all negatively trending ($r_{F5/Overall} = -0.10$, $r_{F5/Abstract} = -0.03$, $r_{F5/News} = -0.09$). Table 1 displays the mean, standard deviation, Pearson's r, and p-values related to each factor when correlated with overall AI differentiation, as well as with specifically abstracts and news headlines only.

Table 1. Sample and correlational statistics for individual factor scores.

Factor	M	SD	Overall AI differentiation (r)	p	Abstracts (r)	p	News (r)	p
F1	28.87	3.41	-0.14	0.1392	0.03	0.7626	-0.21	0.0331
F2	21.03	4.57	-0.14	0.1392	-0.14	0.1372	0.01	0.9377
F3	18.26	3.10	-0.01	0.9518	-0.15	0.1273	0.13	0.1782
F4	13.59	3.32	0.06	0.5504	0.13	0.1713	-0.05	0.5951
F5	11.00	3.55	-0.10	0.3019	-0.03	0.7395	-0.09	0.3322
F6	11.81	3.26	0.06	0.5704	0.05	0.5870	0.02	0.8440

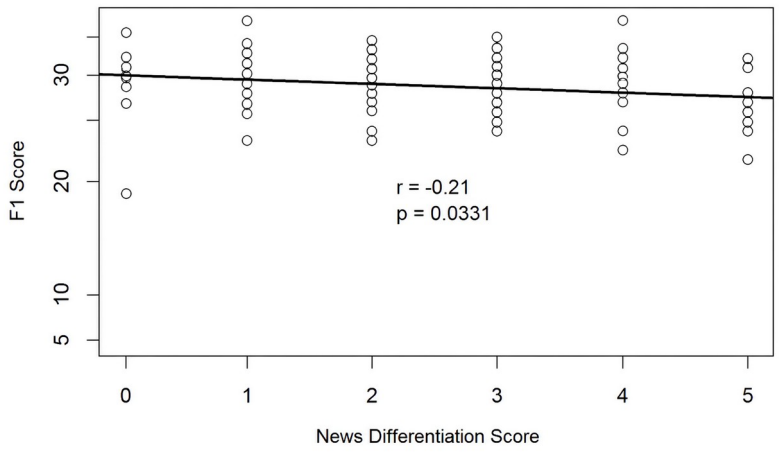

Figure 4. Relationship between F1 scores and AI news headline differentiation scores, as seen in Table 1. Correlational modeling indicates that there is a significantly negative relationship between reflexivity and the ability to identify AI-generated news headlines.

There was a seemingly significant relationship between F1 scores and the accuracy when identifying AI-generated news headlines, which is shown in Figure 4 ($r = -0.21$, $p = 0.0331$). However, because multiple correlational analyses were conducted, the p-values were additionally adjusted according to the Benjamini-Hochberg false discovery rate, and the correlation between F1 scores and AI News Headline differentiation was consequently not found significant. To further investigate whether individual factors obscured by aggregation might predict AI differentiation, individual factor regressions were performed.

3.3 Exploratory analysis

Following peer review, three additional exploratory analyses were performed to investigate whether alternative representations of the MMQ's data were able to predict AI differentiation ability.

3.3.1 *Individual factor regression analysis*

To address the possibility that aggregating the six MMQ factors into good and poor mentalizing composites obscured meaningful individual relationships, a multiple linear regression using Ordinary Least Squares Regression was performed with all factors entered simultaneously as predictors of an individual's overall AI differentiation score. The model, overall, was not statistically significant ($F(6, 100) = 1.645$, $p = 0.142$, $R^2 = 0.092$, adjusted $R^2 = 0.038$), indicating that the MMQ factors together did not reliably predict differentiation performance. Despite the model remaining non-significant, F1(Reflexivity) displayed the largest negative standardized coefficient ($\beta = -.179$, $p = .017$), suggesting a possible unique association after accounting for the remaining MMQ dimensions, and F4 (Relational Discomfort) showed the largest positive coefficient ($\beta = .135$, $p = .063$). Despite this, neither coefficient remained significant following FDR correction. All of the remaining factors were not found significant. The full results of the OLS regression can be found in Table 2, and Figure 5 presents the Beta coefficients and 95% confidence intervals for each factor from the OLS regression.

Table 2. Results of Ordinary Least Squares Regression for all MMQ factors and overall AI differentiation score.

Factor	β	SE	t	p	FDR-adjusted p
(Intercept)	7.101	1.668	4.258	0.000	0.000
F1	-0.179	0.074	-2.423	0.017	0.060
F2	-0.011	0.045	-0.238	0.812	0.812
F3	0.097	0.067	1.461	0.147	0.206
F4	0.135	0.072	1.881	0.063	0.147
F5	-0.094	0.058	-1.612	0.110	0.193
F6	0.083	0.067	1.226	0.223	0.260

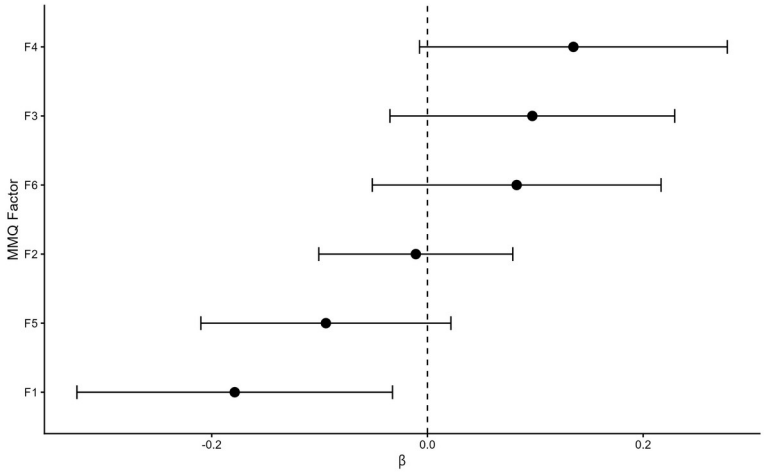

Figure 5. Beta coefficients and 95% confidence intervals for all factors in the Ordinary Least Squares Regression, as seen in Table 2. Correlational modelling initially indicates that F1 (Reflexivity) is negatively correlated with AI differentiation score.

3.3.2 *Principal Component Analysis*

Principal Component Analysis (PCA) of the 6 MMQ factors yielded a two-component solution: PC1 explained 41.4% of the variance and PC2 explained 26.2%, together accounting for 67.6% of total variance. To evaluate whether latent cognitive aspects predicted differentiation outcomes, Pearson correlations were computed between PC1, PC2, and performance scores for the overall AI differentiation task, only News Headline differentiation, and AI Abstract differentiation. The strongest uncorrected effect was observed between PC2 and AI Abstract differentiation ($r = -0.215$, $p = 0.022$), which did not remain significant after correction ($p_{FDR} = .130$).

3.3.3 *Bootstrap analysis*

Bootstrap analyses were also performed for all regression models; however, only the News Headline correlational model demonstrated coefficients which required further interpretation. Bootstrapped confidence intervals (n = 1000) for the News Headline regression indicated that PC2 (95% CI: [0.687, 3.816]) and good mentalizing (95% CI: [−0.597, −0.093]) yielded confidence intervals excluding 0, and were therefore robust. By contrast, PC1 (95% CI: [−0.196, 2.291]) and poor mentalizing (95% CI: [−0.093, 0.210]) performed dissimilarily.

3.3.4 *Cross-validation*

To evaluate external predictive performance, the News Headline regression was assessed using five-fold cross-validation. The model yielded an average cross-validated R^2 of −0.029 (SD = 0.090), indicating that the predictive performance on out-of-sample data was poorer than if a baseline model predicted the mean

outcome. Figure 6 displays the R^2 values for the five-fold cross-validation, as well as the mean cross-validated R^2.

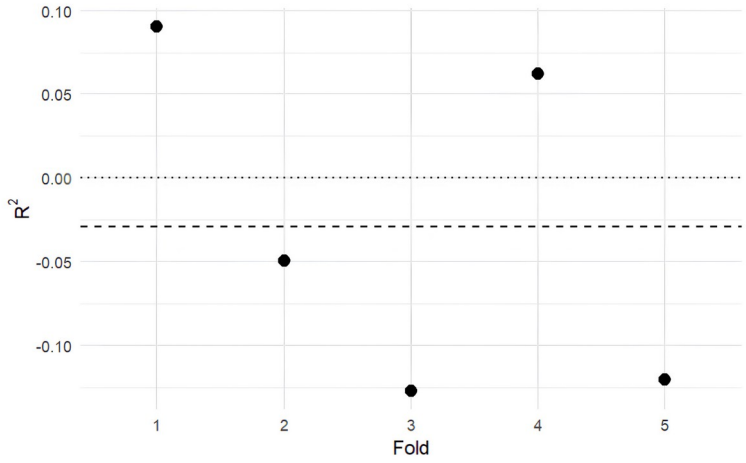

Figure 6. Cross-validated R^2 values for each five folds of the News Headline regression. The dashed line corresponds to the average cross-validated R^2 (−0.029), while the dotted line represents $R^2 = 0$. The negative mean indicates that the model does not generalize well to external data, suggesting overfitting despite stable bootstrap estimates.

3.3.5 *Comparison of three competing regression models*

To determine whether increasingly complex models improved AI News Headline Discrimination, three regression models were compared using AIC and BIC, as explained in Section 2.3.3. The comparison revealed a divergence between AIC and BIC, with AIC favoring Model 3 (the model including all 6 MMQ factors), and BIC selecting Model 1 (The simplest, using only good mentalizing as a predictor). Model 2 (good and poor mentalizing) did not outperform either alternative model under either criterion.

4. Discussion

This study was conducted in order to determine the cognitive factors that affect an individual's ability to discriminate between human and AI writing, with a focus on perceived mentalizing skill and other related latent factors. The findings largely agree with previous literature, while also extending it by evaluating multiple representations of perceived mentalizing systematically. Across all analytic frameworks examined (including correlation, multivariate regression, PCA, and cross-validation) the results consistently failed to identify a robust relationship between self-reported mentalizing and AI differentiation performance. These findings suggest that broad self-reported mentalizing traits, as measured in the present study, provide limited explanation for individual differences in AI differentiation ability under the specific technological conditions examined.

It was hypothesized that, since mentalizing skill is positively correlated with fluid intelligence, it

would also be positively related to AI differentiation skill. Neither good mentalizing (F1-F3) nor poor mentalizing (F4-F6) score composites were significantly associated with any type of AI differentiation, and the individual MMQ factors provided little evidence of meaningful relationships once the probability of false discovery was controlled for. These findings broadly corroborate previous work which found weak or absent relationships between empathy-related factors and AI differentiation (12, 32), while also further suggesting that self-reported cognitive traits of this variety may not substantially contribute to successful AI attribution. Among the initial and exploratory analyses, F1 (which measures Reflexivity) demonstrated the strongest nominal relationship with News Headline differentiation, and also emerged as the largest unique coefficient in the OLS multiple regression model. However, neither finding remained significant following FDR correction, and the News Headline regression demonstrated poor out-of-sample generalizability during five-fold cross-validation despite stable bootstrap estimates.

The PCA of the 6 MMQ factors yielded both a PC1 and PC2, which collectively explained around 67% of the data's variance, but neither component was able to survive FDR correction. This indicates that, although the MMQ factors exhibit a clear latent structure, the structure captured by the PCA does not translate into reliable prediction of differentiation outcomes, nor have any additional predictive value beyond the original factor groupings. Additional resampling analyses provided mixed support for the stability of each individual coefficient. Bootstrap intervals suggested that PC2 and good mentalizing were relatively stable predictors of News Headline differentiation, whereas PC1 and poor mentalizing were not. However, these apparent effects did not generalize in cross-validation, where model performance was below the baseline expectation ($R^2 < 0$).

Nevertheless, bootstrap analysis and cross-validation evaluate different statistical properties, and should therefore not be interpreted interchangeably. While bootstrap resampling assesses the stability of population parameter estimates within repeated samples drawn from the original sample's population, cross-validation conversely evaluates how well a model predicts previously unseen observations. Consequently, a model may exhibit stable coefficient estimates while still demonstrating poor out-of-sample generalizability. Although the estimates for PC2 and good mentalizing were relatively stable, the overall News Headline model demonstrated poor out-of-sample predictive performance. Taken together, these findings suggest that the observed relationships were internally consistent within the sample, but are unlikely to reliably generalize to other datasets.

Model comparison using AIC and BIC further highlighted the lack of a clearly superior explanatory framework. Although AIC slightly favored the most complex model, which included all MMQ factors, BIC preferred the simplest model, and intermediate models showed no consistent favorability. This divergence suggests that improved predictive fit for more complex models did not outweigh the penalty for increased model complexity.

One of the more central findings, perhaps, of the present study was methodological, rather than theoretical. One possible explanation for the absence of significant relationships is that mentalizing is inherently multidimensional, rather than a single unified construct (25, 27). Understanding this, one methodological question explored in the current study was whether aggregating the 6 MMQ factors into broad good and poor mentalizing composites could obscure relationships that might emerge when the factors were analyzed separately or represented using alternative approaches. As such, the present study compared several theoretically motivated representations of perceived mentalizing, including higher-order composites, individual MMQ factors, Principal Components, and multivariate regression models.

Despite these various alternative analytical approaches, no representation consistently demonstrated robust predictive capability, and the data consistently failed to support a predictive signal, suggesting that the predominantly null findings are unlikely to reflect a limitation of any single statistical approach or scoring method. Rather, they suggest that the relationship between perceived mentalizing and AI differentiation, if present, is likely to be considerably weaker than originally hypothesized, and may be unstable or highly sensitive to modeling choices. These findings may also indicate that self-reported mentalizing is not the primary dimension underlying AI differentiation ability in the context of this task.

Taken together, these findings illustrate the value of examining psychological hypotheses across multiple, complementary analytical models. Had the interpretation relied solely on nominally significant correlations, the present study might have concluded that Reflexivity was a meaningful predictor of AI differentiation. However, incorporating the various exploratory analyses demonstrated that this apparent relationship was neither robust nor consistently generalizable, and this apparent relationship was not statistically significant.

4.1 Implications and future directions

The present findings should not be interpreted as evidence that social cognition or mentalizing are unimportant in human interactions with AI. Rather, they suggest that self-reported mentalizing, as operationalized by the MMQ, was not a robust predictor of performance on the present AI differentiation task. This distinction is important because the absence of an association with one operationalization does not preclude the possibility that other aspects of social cognition contribute to judgments about AI-generated content.

Previous research suggests that individuals frequently rely on linguistic and stylistic heuristics—including tone, structure, fluency, and perceived personality—when distinguishing AI-generated from human-authored text, rather than deeper social-cognitive processes (15, 34). Consequently, successful AI differentiation may depend more strongly on linguistic pattern recognition, executive functioning, familiarity with AI-generated language, critical evaluation of writing style, or domain-specific expertise than on stable self-reported sociocognitive traits (10, 12). Future investigations should therefore broaden their focus beyond empathy-related constructs to examine alternative cognitive

mechanisms, including executive functions (e.g., planning, working memory, cognitive flexibility), behavioral measures of Theory of Mind, linguistic sensitivity, and metacognitive reasoning.

Future studies should also evaluate whether alternative measures of social cognition, including behavioral assessments of mentalizing alongside self-report instruments, more accurately predict AI differentiation performance. Larger and more diverse samples, multimodal cognitive assessments, and longitudinal or repeated cross-sectional designs may help determine whether the patterns observed here reflect stable cognitive relationships or context-specific findings.

More fundamentally, future research should seek to establish the construct validity and temporal stability of *AI differentiation* itself. Unlike many psychological tasks that involve relatively stable human behaviors, AI differentiation requires judgments about outputs generated by rapidly evolving algorithmic systems whose linguistic characteristics change across model generations. Consequently, it remains uncertain whether AI differentiation represents a stable psychological construct with enduring cognitive determinants or whether it reflects performance against a continually evolving technological artifact. Addressing this question will require evaluating performance across multiple LLM generations, model families, prompting strategies, and domains of generated content. Establishing the stability of the construct itself is likely to be a prerequisite for identifying enduring cognitive predictors of AI differentiation.

4.2 Limitations

Though the data collected provides insight into the continually evolving intersection of AI discrimination and cognitive markers, it is important to recognize and consider its limitations.

Firstly, gAI and AI technology are rapidly evolving fields: though the most recent ChatGPT model was used in this study, it is likely that new models will be developed and released in the near future, altering the validity of the results presented in this study. Additionally, competitor models such as Google's Gemini and X's Grok were not used in this study, but are also increasing in sophistication without regulation from academia (3). Moreover, the outputs of other LLMs may differ substantially from those of ChatGPT, potentially requiring different cognitive strategies or abilities for successful differentiation.

A more fundamental limitation of the present study concerns the validity of "AI differentiation ability" as a construct itself. Unlike many psychological tasks that measure responses to relatively stable human behaviors or cognitive phenomena, AI differentiation involves identifying outputs generated by rapidly evolving algorithmic systems (11). Though the present study considers AI differentiation ability as a quantifiable human ability for the purposes of empirical investigation, as LLMs continue to change in architecture, training data, and stylistic characteristics, the cognitive processes required for successful differentiation may change likewise. As such, it remains uncertain whether AI Differentiation represents a stable human psychological construct with enduring cognitive

predictors or performance against a continually evolving technological artifact. The present findings should therefore be interpreted within the context of the specific LLM (ChatGPT) and experimental paradigm employed, and future work should further establish the construct validity and temporal stability of AI differentiation as a human ability before drawing broader conclusions about its cognitive determinants.

Additionally, especially as AI usage increases in academia and in the generation of scientific abstracts (8), it is imperative to consider the possibility that the excerpts sourced as "human-generated" may very well have been written with the aid of or completely by gAI. To prevent this from occurring, all abstracts were chosen from peer-reviewed, generally acclaimed scientific journals for each field. Despite this precaution, it is uncertain whether or not the abstracts used were completely free of AI influence, as AI has been shown to be able to slip under peer reviewers' scrutiny (33). To attempt to ensure the same quality of news headlines, excerpts were sourced solely from the New York Times; a credible news source.

Another area where limitations may have been introduced was in the study design. Although the sample size was adequate for detecting moderate associations, the study was likely underpowered to detect small effect sizes. Consequently, the absence of statistically significant relationships should not be interpreted as evidence that small but potentially meaningful associations do not exist. The MMQ, though the most recent and advanced measure for mentalizing that has been validated, relies on self-reported scores, which are likely to be subject to upward social desirability bias and may require individuals to evaluate aspects of themselves over which they do not have complete awareness (25). To capture a more complete and accurate picture of an individual's true mentalizing skill, it may be advisable to use multimodal approaches in the future. Additionally, since the MMQ is a self-report instrument, the present findings concern perceived mentalizing rather than behavioral mentalizing performance (25, 27). As a result, the observed relationships should not be interpreted as evidence that mentalizing itself is unrelated to AI differentiation, but rather, that self-perceived mentalizing was not a predictor in this sample.

Finally, many of the regression, PCA, bootstrap, and cross-validation analyses were exploratory in nature and were performed following the initial primary null findings. Although these analyses were theoretically motivated and incorporated procedures designed to reduce false positives, they should be interpreted as generating hypotheses for future investigation, rather than providing confirmatory evidence. Although alternative statistical representations of the MMQ were explored, the current study does not investigate whether alternative psychologically motivated combinations of the 6 MMQ factors might outperform the predefined good and poor score composites. Such analyses may provide additional insight into the dimensional structure of mentalizing, but are not necessary to support the primary conclusions of the present study.

5. Conclusion

This study found that self-reported mentalizing, as measured by the Multidimensional

Mentalizing Questionnaire (MMQ), was not a robust predictor of participants' ability to distinguish AI-generated from human-authored text under the experimental conditions examined. Across multiple complementary analytical approaches—including composite scores, individual MMQ factors, multivariable regression, principal component analysis, bootstrap resampling, cross-validation, and model comparison—no reliable or generalizable association was identified. These findings suggest that self-reported mentalizing, as operationalized in the present study, provides limited explanatory value for AI differentiation performance.

Importantly, these findings should not be interpreted as evidence that mentalizing, more broadly, plays no role in human evaluation of AI-generated content. Rather, they indicate that this particular self-report operationalization did not consistently predict performance in the present paradigm. Other aspects of social cognition, behavioral measures of mentalizing, executive function, linguistic expertise, or domain-specific reasoning may prove more informative and warrant further investigation.

More fundamentally, the present findings highlight an important conceptual issue regarding the construct of *AI Differentiation* itself. Unlike many psychological tasks that assess responses to relatively stable human behaviors, AI differentiation requires judgments about outputs generated by rapidly evolving algorithmic systems whose linguistic characteristics change across model generations. Consequently, it remains uncertain whether AI differentiation represents a stable psychological construct with enduring cognitive determinants or whether it reflects performance against a temporally specific technological artifact. The present study therefore contributes not only evidence regarding self-reported mentalizing but also emphasizes the need to establish the construct validity, temporal stability, and generalizability of AI differentiation before broader claims regarding its cognitive predictors can be made.

Future research should evaluate AI differentiation across multiple generations of large language models, prompting strategies, and domains of generated content while incorporating both behavioral and self-report measures of social cognition. Such work will help determine whether stable cognitive mechanisms underlie AI differentiation or whether performance is primarily shaped by the continually evolving characteristics of generative AI systems.

Acknowledgements

This research was conducted independently, without external funding or advising. All data was collected using Google Forms and analysed using R. Sections of this work were proofread by faculty at the corresponding author’s school for clarity and simplicity. All experimental designs, interpretations, and conclusions are the author's own.

6. References

1. Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 49, 433–460.

2. Rashid, A.B. & Kausik, M.A.K. (2024) AI Revolutionizing Industries Worldwide: A Comprehensive Overview of Its Diverse Applications. *Hybrid Advances*, 7, Article 100277. https://doi.org/10.1016/j.hybadv.2024.100277

3. Monteith, S., Glenn, T., Geddes, J. R., Whybrow, P. C., Achtyes, E., & Bauer, M. (2024). Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2), 33–35. https://doi.org/10.1192/bjp.2023.136

4. Yousaf M. N. (2025). Practical Considerations and Ethical Implications of Using Artificial Intelligence in Writing Scientific Manuscripts. *ACG case reports journal*, 12(2), e01629. https://doi.org/10.14309/crj.0000000000001629

5. Fleckenstein, J., Meyer, J., Jansen, T., Köller, O., Keller, S. D., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers & Education: Artificial Intelligence*, 6, Article 100209. https://doi.org/10.1016/j.caeai.2024.100209

6. Waltzer, T., Cox, R. L., & Heyman, G. D. (2023). Testing the ability of teachers and students to differentiate between essays generated by ChatGPT and high school students. *Human Behavior and Emerging Technologies,* 2023, Article 1923981. https://doi.org/10.1155/2023/1923981

7. Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). *AI vs. human: Differentiation analysis of scientific content generation* (arXiv Preprint No. arXiv:2301.10416). arXiv. https://doi.org/10.48550/arXiv.2301.10416

8. Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, 2(3), 100068. https://doi.org/10.1016/j.rmal.2023.100068

9. Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). *Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews*. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 29575–29620). *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v235/liang24b.html

10. Hadan, H., Wang, D. M., Mogavi, R. H., Tu, J., Zhang-Kennedy, L., & Nacke, L. E. (2024). The great AI witch hunt: Reviewers' perception and (Mis) conception of generative AI in research writing. *Computers in Human Behavior: Artificial Humans*, 2(2), 100095. https://doi.org/10.1016/j.chbah.2024.100095

11. Yang, B.; Qu, J. From Origins to Future: The Evolution and Prospects of Artificial Intelligence in the Reasoning Era. *Journal of International Economy and Global Governance* 2025, 2 (2), 84-96. https://doi.org/10.12414/jiegg.250453

12. Chein, J. M., Martinez, S. A., & Barone, A. R. (2024). Human intelligence can safeguard against artificial intelligence: individual differences in the discernment of human from AI texts. *Scientific reports*, 14(1), 25989. https://doi.org/10.1038/s41598-024-76218-y

13. Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ digital medicine*, 6(1), 75. https://doi.org/10.1038/s41746-023-00819-6

14. Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., Çakir, D. C., & Gerjets, P. (2022). *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 First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)* (pp. 60–61). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.in2writing-1.8

15. Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). *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)* (pp. 7282–7296). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.565

16. Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America*, 120(11), e2208839120. https://doi.org/10.1073/pnas.2208839120

17. Fiedler, A., & Döpke, J. (2025). Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses. *International Review of Economics Education*, 49, 100321. https://doi.org/10.1016/j.iree.2025.100321

18. Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). *Automatic detection of generated text is easiest when humans are fooled.* In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1808–1822). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.164

19. Dik, S., & Erdem, O. (2026). Assessing GPT-Zero's Accuracy in Identifying AI vs. Human-Written Essays. *Proceedings of International Mathematical Sciences,* 7(2), 54-58. https://doi.org/10.47086/pims.1762132

20. Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

21. Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023). Real or Fake Text?: Investigating Human Ability to Detect Boundaries between Human-Written and Machine-Generated Text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11), 12763–12771. https://doi.org/10.1609/aaai.v37i11.26501

22. Moravec, V., Hynek, N., Skare, M., Gavurova, B., & Kubak, M. (2024). Human or machine? The perception of artificial intelligence in journalism, its socio-economic conditions, and technological developments toward the digital future. *Technological Forecasting and Social Change*, 200(0040-1625), 123162. https://doi.org/10.1016/j.techfore.2023.123162

23. Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemans, J. (2018). Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts*, 12(2), 177–192. https://doi.org/10.1037/aca0000136

24. Weinstein, N. Y., Whitmore, L. B., & Mills, K. L. (2022). Individual differences in Mentalizing tendencies. *Collabra: Psychology*, 8(1), Article 37602. https://doi.org/10.1525/collabra.37602

25. Gori, A., Arcioni, A., Topino, E., Craparo, G., & Lauro Grotto, R. (2021). Development of a New Measure for Assessing Mentalizing: The Multidimensional Mentalizing Questionnaire (MMQ). *Journal of personalized medicine*, 11(4), 305. https://doi.org/10.3390/jpm11040305

26. Fonagy, P., Gergely, G., & Jurist, E. L. (Eds.). (2002). *Affect regulation, mentalization and the development of the self* (1st ed.). Routledge. https://doi.org/10.4324/9780429471643

27. Vera Cruz, G., Rochat, L., Liberacka-Dwojak, M., Wiłkość-Dębczyńska, M., Khan, R., & Khazaal, Y. (2024). Validation of the english version of the Multidimensional Mentalizing

Questionnaire (MMQ). *BMC psychology*, 12(1), 344. https://doi.org/10.1186/s40359-024-01837-z

28. "R/SurveyExchange." *Reddit*, www.reddit.com/r/SurveyExchange/

29. "R/SampleSize." *Reddit*, 2012, www.reddit.com/r/SampleSize/

30. OpenAI. (2026). ChatGPT (GPT-5.2) [Large language model]. https://chat.openai.com/

31. Zhu, T., Weissburg, I., Zhang, K., & Wang, W. Y. (2025). *Human bias in the face of AI: Examining human judgment against text labeled as AI generated.* In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 25907–25914). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.1329

32. Hitsuwari, J., Ueda, Y., & Nomura, M. (2022). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior,* 139, 107502. https://doi.org/10.1016/j.chb.2022.107502

33. Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). *Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews.* In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 29575–29620). *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v235/liang24b.html

34. Markowitz, D. M., Hancock, J. T., & Bailenson, J. N. (2023). Linguistic Markers of Inherently False AI Communication and Intentionally False Human Communication: Evidence From Hotel Reviews. *Journal of Language and Social Psychology*, 43(1), 63-82. https://doi.org/10.1177/0261927X231200201

35. Zwebner, Y., Sellier, A. L., Rosenfeld, N., Goldenberg, J., & Mayo, R. (2017). We look like our names: The manifestation of name stereotypes in facial appearance. *Journal of personality and social psychology*, 112(4), 527–554. https://doi.org/10.1037/pspa0000076

36. Yamaguchi, K., & Kempf, A. (2026). Encrypted Qubits Can Be Cloned. *Physical review letters*, 136(1), 010801. https://doi.org/10.1103/y4y1-1ll6

37. Li, Z., Liu, Z., Wang, Z., Deng, Y., Yang, S., Chen, J., Zeng, Q., Zhong, Y., Yang, H., Xiong, Z., Tian, X., Li, G., Chen, Y., Jing, H., Ho, J. S., & Qiu, C.-W. (2026). *Body sensor*

networks based on flexible topological clothing. *Nature Electronics, 9*(1), 59–68. https://doi.org/10.1038/s41928-025-01516-w

38. Sammarco, I., Krtilová, E., Slovák, M., & Lafon Placette, C. (2025). Reversibility of sex changes in the plant kingdom: more important than we thought?. *Biological reviews of the Cambridge Philosophical Society*, 100(6), 2199–2216. https://doi.org/10.1111/brv.70043

39. Kim, J., & Munster, P. N. (2025). Estrogens and breast cancer. *Annals of oncology : official journal of the European Society for Medical Oncology*, 36(2), 134–148. https://doi.org/10.1016/j.annonc.2024.10.824

40. Bradsher, K. (2025b, December 23). China Delays Plans for Mass Production of Self-Driving Cars After Accident. *The New York Times*. https://www.nytimes.com/2025/12/23/business/china-autonomous-cars-driving.html

41. Jacobs, J., & Sisario, B. (2025, December 24). Sean "Diddy" Combs's Lawyers File Appeal, Arguing His Sentence Was Unjust. *The New York Times*. https://www.nytimes.com/2025/12/23/arts/music/sean-combs-diddy-appeal-sentence.html

42. Ortiz, A., Smialek, J., Le Stradic, S., Jones, J., & Kwai, I. (2026, January 6). *Winter weather snarls travel across Northern Europe*. *The New York Times*. https://www.nytimes.com/2026/01/06/world/europe/amsterdam-paris-travel-delays-weather.html

43. Audra. (2026, January 8). Minimum Wage Rises in Some States as Workers Struggle With Basic Costs. *The New York Times*. https://www.nytimes.com/2026/01/08/us/minimum-wage-increases.html

44. Albeck-Ripka, L. (2025, November 13). Canada-U.S. Travel Drops for 10th Month Amid Trump Tariff Tensions. *The New York Times*. https://www.nytimes.com/2025/11/13/us/canada-us-travel-decline-flights.html

Appendix 1. Excerpts used in AI differentiation task

Table A1. Human and AI-generated abstracts

Discipline	Human	gAI
Psychology	Research demonstrates that facial appearance affects social perceptions. The current research investigates the reverse possibility: Can social perceptions influence facial appearance? We examine a social tag	We examined whether culturally shared name stereotypes are reflected in facial appearance, thereby enabling observers and computational algorithms to match persons' names to their faces at

	that is associated with us early in life— our given name. The hypothesis is that name stereotypes can be manifested in facial appearance, producing a face-name matching effect, whereby both a social perceiver and a computer are able to accurately match a person's name to his or her face. In 8 studies we demonstrate the existence of this effect, as participants examining an unfamiliar face accurately select the person's true name from a list of several names, significantly above chance level. We replicate the effect in 2 countries and find that it extends beyond the limits of socioeconomic cues. We also find the effect using a computer-based paradigm and 94,000 faces. In our exploration of the underlying mechanism, we show that existing name stereotypes produce the effect, as its occurrence is culture-dependent. A self-fulfilling prophecy seems to be at work, as initial evidence shows that facial appearance regions that are controlled by the individual (e.g., hairstyle) are sufficient to produce the effect, and socially using one's given name is necessary to generate the effect. Together, these studies suggest that facial appearance represents social expectations of how a person with a specific name should look. In this way a social tag may influence one's facial appearance (35).	rates exceeding chance. Across eight studies, participants viewed unfamiliar adult faces and selected each face's true given name from a set of alternatives. Consistently, human perceivers identified correct name–face pairings above chance, revealing a reliable face-name matching effect that generalized across two national samples (Israel and France) and held after controlling for socioeconomic cues. Complementing human judgment data, a machine-based paradigm employing a large face dataset (~94,000 images) replicated the above-chance matching pattern, suggesting systematic associations between facial appearance and names at scale. Follow-up analyses implicated existing name stereotypes as a driving mechanism: the face–name matching effect diminished when cultural congruence was absent, and elements of appearance under voluntary control (e.g., hairstyle) were sufficient to sustain the effect. These results support a model in which early-acquired social tags, such as given names, are internalized and gradually manifest in appearance, producing consistent cultural expectations about how individuals with specific names should look. Consequently, facial appearance may encode socially constructed identity cues well beyond traditional demographic markers.
Physics	We show that encrypted cloning of unknown quantum states is possible. Any number of encrypted clones of a qubit can be created through a unitary transformation, and each of the encrypted clones can be decrypted through a unitary transformation. The decryption of an encrypted clone consumes the decryption key; i.e., only one decryption is possible, in agreement with the no-cloning theorem. Encrypted cloning represents a new paradigm that provides a form of redundancy, parallelism, or scalability where direct duplication is forbidden by the no-cloning theorem. For example, a possible application of encrypted cloning is to enable encrypted quantum multicloud storage (36).	The no-cloning theorem prohibits the creation of identical copies of an unknown quantum state, severely limiting strategies for redundancy and backup in quantum information processing. Here we present a novel scheme that produces encrypted copies of an arbitrary qubit under unitary evolution, thereby enabling the distribution of encrypted versions of quantum data while preserving the integrity of quantum mechanics. The encoding transforms an unknown state $\|\psi\rangle$ into nnn signal qubits entangled with ancillas forming the decryption key, such that each signal qubit alone remains maximally mixed. A dedicated decryption unitary acting on a chosen signal qubit and all ancillas yields exact state recovery at that site only, consuming the decryption key in the process and prohibiting further extractions. We analyze the resource scaling of the protocol, showing that gate counts increase linearly with the number of encrypted copies, and we discuss the constraints that enforce a single usable decryption per protocol instance. This work highlights a mechanism for redundant encrypted quantum data storage compatible with current quantum hardware and lays a foundation for secure,

		distributed quantum information architectures.
Engineering	Body sensor networks wirelessly interconnect multiple on-body sensors using metamaterials that are capable of supporting microwave near-field or surface-wave propagations. However, the design of such networks is typically restricted to one-dimensional unit-cell structures. Topological metamaterials are often used in photonics applications such as lasers and photon sources, but their integration with biological systems remain limited due to low flexibility, high bending loss and high energy dissipation in biological environments. Here we report flexible topological metamaterial clothing that can provide robust biosensing networks on the human body. The approach is based on two-dimensional topological modules fabricated from thin metallic conductive textiles. The resulting topological edge states improve on-body signal transmission by over three orders of magnitude (more than 30 dB) compared with conventional radiative networks, and can maintain performance under various bending angles. The modular design allows reconfiguration by varying the combination of topological phase modules. We show that the topological clothing with interconnected biosensors, and enhanced with machine learning algorithms, can monitor vital signs during exercise with an over two orders of magnitude improvement in signal-to-noise ratio and a threefold increase in accuracy compared with a system without topological clothing (37).	Flexible topological clothing offers a new paradigm for embedded body sensor networks by combining engineered metamaterial textiles with scalable communications infrastructure. In this study, we design and fabricate conductive textile modules that support topologically protected electromagnetic edge states enabling robust on-body signal propagation. Traditional wireless networks on the body suffer from high loss, poor signal confinement, and motion sensitivity; in contrast, the topological clothing leverages planar valley topological models realized in metallic textile patterns to guide microwave near-field or surface-wave propagation with minimal dispersion and defect susceptibility. Experimental characterization confirms propagation enhancements greater than three orders of magnitude compared with conventional textile networks, with consistent performance under significant fabric deformation and wearer motion. The modular construction enables reconfiguration of topological phases and edge state counts, providing network designers with unprecedented control over channel layout and sensor allocation. Integrated sensor arrays distributed across the garment communicate via these tailored propagation channels, supporting continuous multi-location physiological monitoring. When interfaced with advanced data analytics, the topologically enabled system demonstrates markedly improved vital-sign signal integrity and resilience against motion artifacts, supporting accurate monitoring during exercise and daily activities. These results substantiate the feasibility of topological metamaterial clothing as an engineering foundation for next-generation wearable networks that marry mechanical flexibility with electromagnetic performance rarely attained in textile electronics.
Biology	Compared to animals, plants show a wide range of reproductive strategies with different degrees of sex separation (e.g. dioecy, monoecy, hermaphroditism). While sex expression was previously thought to be genetically determined and fixed in plants, accumulating evidence suggests that sex expression can change reversibly even within one generation (sex changes), involving non-genetic factors (i.e. environment and epigenetics). In addition, recent work suggests that sex determination itself relies on epigenetic factors. Therefore, in this review, we propose that the border between sex changes and the apparently “fixed” determination of sexes is less clear than previously thought, as they rely on similar	Sexual systems in flowering plants span a continuum from hermaphroditism to dioecy, with significant evolutionary transitions documented among these states. Historically treated as genetically fixed entities, sexual phenotypes are now increasingly recognized as labile and reversible within individuals. This review synthesizes recent findings on the frequency and ecological distribution of within-generation sex changes, emphasizing their occurrence across diverse taxa and sexual polymorphisms. We propose that such reversibility, mediated by the interplay of environmental triggers, phytohormonal signals, and epigenetic mechanisms, may have profound evolutionary consequences. In

	mechanisms, in particular epigenetics. Specifically, we propose that within-generation sex changes may facilitate evolutionary transitions between different degrees of sex separation via the assimilation of epimutations into genetic mutations. We then evaluate the (mal)adaptive potential of sex changes. We conclude that in the face of global environmental changes, sex changes may follow a bet-hedging evolutionary strategy, that is a heritable ability to reverse sexes. Sexual bet-hedging with an epigenetic basis (via stochastic epimutations) may help plants alleviate the deleterious consequences of climate change (38).	particular, the capacity for sex change may act as a facilitator of evolutionary sex shifts, allowing populations to traverse sexual system space more readily than predicted by models predicated on fixed genetic determination. We evaluate the adaptive potential of sex reversibility in the context of fluctuating environmental conditions and mating limitations, suggesting that a diversified bet-hedging strategy may enhance reproductive success and buffer populations against climate-driven stressors. Conversely, we also examine cases where sex reversals lead to reduced fitness, highlighting that the evolutionary outcomes of this plasticity depend on ecological context and species-specific life histories. We conclude by identifying key empirical gaps and recommending future studies aimed at quantifying the prevalence, mechanisms, and evolutionary significance of sex reversibility in plants.
Medicine	Estrogens have been associated with an increase in breast cancer risk. Yet emerging clinical and experimental evidence points to progestogens [endogenous progesterone or synthetic progesterone (progestin)] as the primary hormonal driver underlying seemingly estrogen-associated breast cancer risk. Estrogens may contribute to breast cancer risk indirectly by induction of the progesterone receptor and thus amplifying progesterone signaling. Large studies of hormonal contraceptives suggest that the small increase in breast cancer risk from hormonal contraceptives is mainly attributable to progestins, not estrogens. Estrogen-plus-progestin hormone replacement therapy (HRT) has consistently shown an increase in breast cancer risk among postmenopausal women, whereas estrogen-alone HRT has little impact on breast cancer risk in naturally or surgically menopausal women. In particular, the long-term follow-up of the Women's Health Initiative (WHI) randomized trials suggests a benefit of estrogen alone. Recent data further indicate that endogenously elevated estrogen during assisted reproductive technology (ART) exhibits little adverse effect on or potentially a reduction in breast cancer risk and recurrence. Also, accumulating evidence suggests that inhibition of progesterone signaling is a critical mechanism underlying the risk-reducing and therapeutic effects of antiestrogens. Estrogen HRT has shown an array of proven benefits, including ameliorating menopausal symptoms and improving bone health. Collective evidence thus suggests that estrogen HRT is likely to offer health benefits to perimenopausal or postmenopausal women, including breast cancer survivors, as well as young BRCA1/2	The role of estrogens in breast cancer is complex, encompassing receptor-dependent proliferative effects and broader systemic influences on tumor microenvironment and host metabolism. Estrogens exert their principal actions via estrogen receptor signaling, with ERα-mediated transcription driving cell cycle progression, anti-apoptotic pathways, and gene networks that support neoplastic growth in hormone receptor-positive breast tissues. Physiologically elevated estrogen levels, including those associated with reproductive history or adipose tissue aromatase activity post-menopause, contribute to a pro-mitogenic milieu that fosters clonal expansion of initiated cells. Beyond classical genomic pathways, estrogen metabolism generates bioactive metabolites capable of inducing oxidative DNA damage and epigenetic alterations, further implicating estrogen biology in pathogenesis. Although estrogens are often discussed in relation to breast cancer risk, recent evidence underscores that progestogens may modulate estrogen-associated risk by enhancing progesterone receptor expression and downstream signaling. Clinically, this interplay informs hormone replacement therapy strategies, where estrogen-alone regimens show differing risk profiles compared with combined therapies. Endocrine therapies targeting estrogen signaling, such as aromatase inhibitors, SERMs, and selective estrogen receptor degraders, remain foundational in managing hormone receptor-positive disease and reducing recurrence risk. This review synthesizes mechanistic insights with translational evidence, highlighting how estrogens interact with receptor pathways, metabolic context, and systemic factors to

	carriers with prophylactic oophorectomy for ovarian cancer prevention (39).	shape breast cancer risk and inform targeted interventions.

Note. The order of gAI and human excerpts was randomized when presented to participants.

Table A2. Human and AI-generated News Headlines

Human	gAI
China Delays Plans for Mass Production of Self-Driving Cars After Accident (40)	Fatal Accident Forces China to Reassess Autonomous Vehicle Timetable, Regulators Tighten Approvals
Sean Combs's Lawyers File Appeal, Arguing His Sentence Was Unjust (41)	Combs's Legal Team Asks Appeals Court to Overturn Sentence, Calling 50-Month Term Unjust
Winter Weather Snarls Travel Across Northern Europe (42)	Snow, Ice and Fierce Winds Leave Thousands Stranded from Amsterdam to Paris
Minimum Wage Rises in Some States as Workers Struggle With Basic Costs (43)	Minimum Wage Increases Hit 22 States as Workers Grapple With Soaring Living Costs
Canadian Travel to the U.S. Declines for 10th Straight Month (44)	Economic Ripples From Canada as U.S. Bound Travel by Canadians Drops for Tenth Consecutive Month

Note. The order of gAI and Human excerpts was randomized when presented to participants.

Appendix 2. Multidimensional Mentalizing Questionnaire (25)

Please read the following statements carefully and indicate how well they describe you by choosing one of the five response options below. There are no right or wrong answers, so we ask you to honestly choose the answer that you think is most appropriate for you.

Strongly disagree	*Disagree*	*Neither agree nor disagree*	*Agree*	*Strongly agree*
1	**2**	**3**	**4**	**5**

1. I often try to explain what is happening to me	1	2	3	4	5
2. I am an impulsive person	1	2	3	4	5
3. I sometimes experience mood swings I can't control	1	2	3	4	5
4. I'm able to get the deepest aspects of people around me	1	2	3	4	5

5. I can tune in other people's mental states	1	2	3	4	5
6. Understanding what others feel is crucial in understanding their actions	1	2	3	4	5
7. I sometimes feel like I am losing control of my emotions	1	2	3	4	5
8. I am able to reflect on my behaviours	1	2	3	4	5
9. Relationships with other people prevent me from being myself	1	2	3	4	5
10. I'm interested in understanding my mental processes	1	2	3	4	5
11. I can tolerate frustrations of daily life	1	2	3	4	5
12. Others don't understand me	1	2	3	4	5
13. It's better to beware of others	1	2	3	4	5
14. I'm able to empathize with others when they tell me something	1	2	3	4	5
15. I am afraid to open up with other people	1	2	3	4	5
16. I ponder over what happens to me	1	2	3	4	5
17. I find beneficial to analyse my behaviour	1	2	3	4	5
18. I often think about why things happen	1	2	3	4	5
19. For me things are either white or black	1	2	3	4	5
20. I don't trust others	1	2	3	4	5
21. I am sensitive to what happens to others	1	2	3	4	5
22. I can usually adapt myself to different contexts with no difficulties	1	2	3	4	5
23. It happens to me to have conflicting emotions	1	2	3	4	5
24. I am able to sort out difficult problems when life presents those to me	1	2	3	4	5
25. I am able to bear the emotional load of stressful situations	1	2	3	4	5
26. When I feel an intense emotion, I can control it	1	2	3	4	5

27. People abandon me	1	2	3	4	5
28. I can easily attune to other people's thinking	1	2	3	4	5
29. It's better to beware of strangers	1	2	3	4	5
30. I am able to cope with difficult situations	1	2	3	4	5
31. I am a thoughtful person	1	2	3	4	5
32. I'm keen on understanding why certain things happen to me	1	2	3	4	5
33. Some people are the cause of my problems	1	2	3	4	5