Humanities & Social Sciences Communications



COMMENT

https://doi.org/10.1057/s41599-024-04002-4 OPEN

Check for updates

# Human cognition and AI-generated texts: ethics in educational settings

Ceymi Doenyas[1]✉

As artificial intelligence (AI) tools refine their ability to produce increasingly coherent and fluent texts, the ethical implications of their utilization in educational contexts warrant discussion. Upon briefly overviewing recently emerging evidence on AI-generated and human-written texts, I note that detecting AI-generated texts may be difficult not only for humans but also for technological detectors and classifiers. Accordingly, for the investigation of the relationship between human cognition and the recent developments in AI language generators, I suggest that the main focus should be on educating students on using AI only in ways that are conducive to higher-level thinking and learning, more so than how students' AI-generated texts can be better detected. To achieve this, educational curricula should adapt to today's AI revolution by starting to teach students how to utilize AI to support their analytical and critical thinking skills, rather than using AI tools to do their work for them and thereby letting such tools erode these qualities. For this to be possible, students should be motivated to use AI only to support their thinking and learning, which can be possible through the creation of an honor code specific to the ethical utilization of AI in educational contexts. As this AI-specific honor code does not yet exist, I outline ways based on previous honor code research for this code to be created and such a culture of AI ethics to be implemented. In sum, I recommend ethical and practical measures that can facilitate a healthy AI-human partnership in educational settings.

[1] Yıldız Technical University, Department of Guidance and Psychological Counseling, İstanbul, Türkiye. ✉email: ceymidoenyas@alumni.princeton.edu

## Human cognition and AI-generated texts: ethics in educational settings

Today's world is marked by an artificial intelligence (AI) revolution in which students are increasingly using AI writing tools. Some students use these tools to get help on revising and improving their writing, but, unfortunately, other students have AI write their texts for them. Since it is neither possible nor advisable to ban AI tools at schools, ethical values about their appropriate utilization should be taught to avoid such cases happening in the future of education and academia. This is the focus of the present paper, where I discuss the reasons and possible ways for creating a culture of ethical AI utilization in educational contexts.

Humans have evolved to effectively respond to and utilize stimuli around them to increase their chances of survival. While the Darwinian evolutionary perspective focused on physical survival, survival in today's conditions also encompasses success in an ever-changing, increasingly technological world. Considering that definitions of intelligence include the ability to adapt, one mark of human intelligence can be said to be the ability to adapt to the current world where AI is gaining more and more prominence. AI has a multitude of applications in different domains as well as being approached, analyzed, and discussed from various perspectives ranging from systems engineering to social sciences. Here, I approach it from the perspective of human cognition. Upon overviewing how humans perceive AI, as well as their confidence in their ability to detect AI-generated text, I discuss the implications of such human perceptions of AI for learning and teaching contexts. Specifically, I propose the need for a shift of focus from whether or not AI-generated text can be detected to educating students on how to use AI in ways that support their higher-level thinking and learning, which I believe should be the main focus of how human cognition relates to the recent developments in AI language generators.

**Thinking about how machines think**. In 1950, Alan M. Turing asked, "Can machines think?" and then proposed The Imitation Game as a test for this query. In this game, there is a man, a woman, and an interrogator, and the interrogator needs to determine the gender of each person through questions. Turing then asks what would happen if a machine took the part of one of the individuals in this game. He replaces his original question of "Can machines think?" with whether or not the interrogator will make the same percentage of mistakes when the game is played with a machine as opposed to a human man and woman (Turing, 1950). The Turing Test has been the topic of innumerable investigations and discussions. One comment on it that particularly resonates with my cognitive approach toward AI-human interactions comes from Joseph Fink: "It turns out this whole time that the Turing Test was the wrong way to think about it. Thinking a chatbot is alive is not a test of how good the chatbot is, but of your own ability to think of other human beings as real and complete people" (Fink, 2024). He considers AI or chatbots as novel agents for social interaction and believes that we apply the same mentalizing processes we apply to humans when thinking about chatbots. Therefore, he deduces that the Turing Test is more of a measure of how good our ability is to mentalize about other people than the chatbot's ability to fool us into believing it is human. This interpretation fits well with my suggestion about evaluating how humans interact with AI through a focus on human cognition. Likewise, Watt (1996) has previously contended that "a system passes the Turing test if the observer believes that it [the system] has a mind." He also points out the natural human tendency to assign mental states to others, a tendency that gives an evolutionary survival advantage in complex societies (Watt, 1996).

Another evolutionary advantage besides attributing mental states to others is the ability to evaluate the accuracy of one's own mind. When navigating the world, we engage our cognitive processes to perceive and interpret the stimuli in our environment and to decide on an appropriate action, which we then execute. To increase the probability of performing actions that lead to success, we evaluate the contents and outputs of our cognition for accuracy. This evaluation is called metacognition, or "cognition about cognition," referring to any cognitive process about another cognitive process (Kralik et al., 2018). Metacognition involves monitoring and controlling one's own cognitive processes, which, when applied to others, is called mentalizing (Frith, 2012). Interestingly, and rather surprisingly, humans are said to have limited access to underlying processes used to understand themselves and others, and therefore they can be inaccurate when reporting their own and others' intentions (Frith, 2012). When we try performing metacognitive thinking and mentalizing, we often use intuition or mental shortcuts (heuristics). As aptly noted by the leading scientists in this topic, Tversky and Kahneman (1974), "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors." Nowadays, with the 'birth' of AI, we are tasked with making judgments about the sources of texts we receive in online interactions and academic settings. In these contexts, we need to evaluate whether or not we are talking with an actual person on online platforms or if the assignments handed in by students are written by them or by an AI language generator. Since such evaluations are relevant for the utilization of AI in educational contexts, some recently emerging evidence about our ability to detect AI-generated texts and the heuristic strategies used in doing so is worth noting.

**Accuracy and metacognition about AI-generated content**. Humans were not completely deceived by AI-generated poetry when it was random and was not refined, but when the best poem produced by AI was chosen, humans failed to reliably detect it as GPT-produced, even when incentives were provided for correct detection (Köbis and Mossink, 2021). Metacognition about this ability to detect AI-generated text was explored in another incentivized task paradigm that utilized deepfakes, which are overly realistic manipulations done on video and sound content. People could not reliably detect deepfakes, and their accuracy was not reliably improved after providing financial incentives or after raising awareness by reading a text about deepfakes. The researchers found that people were overconfident in their relatively low detection abilities (Köbis et al., 2021). Carrying these evaluations to the social realm, when state-of-the-art AI language models generated self-presentations in dating, hospitality, and professional contexts, people were unable to detect that these were generated by AI (Jakesch et al., 2023). These findings suggest that people are not very good at detecting high-quality or refined AI-generated content, and they tend to overestimate their ability to do so.

**Heuristic strategies when evaluating text sources as AI vs. human**. Jakesch et al. (2023) analyzed the heuristics humans use when trying to distinguish AI-generated from human-written verbal self-presentations. The first two sentences of some of the prompts are listed here for clarification purposes. For the professional context, one human prompt starts with "I have 19 years of journalism experience. My work has appeared in daily and weekly newspapers, international trade magazines and

textbooks," while one GPT-generated prompt begins as "My name is Gary Stauch, and I have been in the computer and electronics business for over 30 years. I have an A.S. in electronics, a B.S. in computer science and I am a registered professional engineer in Texas." For the dating context, one human prompt ends with the sentence "I enjoy being fit, active, and healthy, though I do eat ice cream for dinner on occasion," while a GPT-generated and GPT-optimized prompt ends with "I'm easy going and down to earth, I know what I want in life and am working towards my goals. message me if you want to know more :)" (Jakesch et al., 2023).

Participants considered self-presentations with nonsensical content and repetitive content as AI-generated, which helped them detect AI-generated content with these characteristics. Therefore, not all used heuristics were flawed, in line with the cognitive psychology literature (e.g., Tversky and Kahneman, 1974). Yet, despite the accuracy of some, most used heuristics were flawed. While participants considered grammatical issues to indicate
AI-generated content, grammatically flawed self-presentations in the study were less likely to be generated by AI. Participants thought the usage of long words or rare bigrams indicated AI generation, while most self-presentations that used long words or rare bigrams were written by humans. Participants evaluated the use of first-person speech, content about family, the inclusion of authentic or spontaneous words, focus on past events, and longer self-presentations as more human, which were, in fact, not significantly associated with either human or AI-generated language in this study's data. No gender or ethnic group was better than others in detecting AI-generated self-presentations. Interestingly, university graduates performed 1% worse than those with no university degree, and self-reported technical knowledge did not correlate with more accurate ratings (Jakesch et al., 2023). This can relate to the overconfidence coming from a university degree or technical knowledge that can make it more likely for individuals to trust their heuristic judgments. The time taken for the judgment not predicting higher judgment accuracy is in line with the automatic and quick nature of heuristics, implying that this was the likely decision-making process involved here as opposed to more deliberate, time-consuming processes (see a discussion of System 1 and System 2 thinking by Kahneman (2013)).

In this study, a language model had a 58.1% accuracy rate in predicting participants' evaluations of AI-generated language, suggesting that the heuristics people use for these evaluations are somewhat predictable (Jakesch et al., 2023). The authors then created texts that would circumvent these heuristics and appear to be written by humans. The participants rated these optimized AI-generated self-presentations as more human than the actual human-written self-presentations. Jakesch et al. (2023) interpreted this as heuristics making "human judgment of AI-generated language predictable and manipulable, allowing AI systems to produce text perceived as more human than human."

By extending findings from the decision-making literature to the use of heuristics in deciding whether a text is generated by AI or not, I suggest that the human inclination for heuristic use is not specific to making decisions about oneself and other humans but also applies to domains involving AI, implying a conserved cognitive strategy in personal, social, and technological contexts. While presenting the recent findings about human cognition and AI, I acknowledge that this discussion of recently emerging evidence is not conclusive. It is a possibility that future results may refute the discussed findings or determine the conditions in which they apply and those in which they do not. It is also likely that future studies can reveal cross-cultural differences in AI-human cognitive interactions. These are exciting future avenues to look forward to, which are not so far ahead in time, given the fast pace of studies being conducted and published on human-AI interactions.

**AI-generated vs. human-written text**. Are AI-generated and human-written texts obviously different from each other, and are we unable to distinguish between them because we use flawed heuristics? Or are these texts perhaps hardly distinguishable?

In terms of scientific content, although AI could create scientific content with similar accuracy to human-written content, these two differed in terms of depth and overall quality, and AI-generated text was more likely to include factual errors, leading the authors to infer a "writing style" gap between AI and human-written scientific text (Ma et al., 2023). AI-generated scientific text, especially the scientific abstract, only contained generalities and missed valuable insights and had a low external consistency with real scientific knowledge (Ma et al., 2023). Setting scientific accuracy aside, these other minor differences may not be salient enough to be reliably detected by humans. Additionally, language barriers or previous academic experience can lead individuals to create scientific texts possessing these very qualities that are said to distinguish AI-generated from human-written text.

Not just a difficulty for humans, the detection of AI-generated text is also a challenge for technological detection methods. AI tools using large language models such as ChatGPT can create original and coherent text that can go undetected by both technological tools and trained academics (Perkins, 2023). Technically, when a light paraphraser is implemented on top of the generative text model, it is reported to be able to break many detectors and classifiers (Sadasivan et al., 2023), suggesting that this already tough task is being made even more difficult with strategies such as applying paraphrasers after the initial text generation. This led the authors to conclude that "for a sufficiently good language model, even the best-possible detector can only perform marginally better than a random classifier" (Sadasivan et al., 2023), where it may never be possible to create a detector that can reliably distinguish AI-generated text. Similarly, Ethan Mollick of Wharton School notes that even when small changes are made to AI-generated texts, such as asking the AI to revise its own text, this process and any utilization of iterations can fool the detectors. He also notes that there are false positives with machine detection of AI-generated text (Mollick, LinkedIn post, April, 2023).

These remarks suggest that it is very difficult not only for humans but also for technological systems to detect AI-generated text. Though difficult, this task is not impossible, with research groups starting to use probability functions (Mitchell et al., 2023), smaller language models (Mireshghallah et al., 2024), and white-box detection methods instead of the black-box detecting methods that are becoming less effective as language models improve (Tang et al., 2024) to better detect AI-generated text. One group of researchers showed that having the AI-generated text detector use semantic retrieval increases the detector's robustness in the face of paraphrasing (Krishna et al., 2024). This suggests that teaching computers to operate more like the human mind does, such as by using semantic retrieval, can beat paraphrases, reiterations, and revisions of AI-generated texts. As these strategies are also likely being devised by human minds who want to avoid the detection of their AI-generated text, teaching the AI-generated text detector to operate more like the human mind seems like a good starting point in developing detectors that are robust and sensitive to strategies for avoiding detection.

**Recommendations for educational institutions**. Understanding if teachers can detect AI-generated text or not and developing

software that can effectively do so is an important mission. However, would it not be a better teaching practice to inoculate students with ethical values that will guide them to use AI ethically and declare it when they do? In terms of ethical considerations for AI utilization in academic contexts, Perkins (2023) contends that plagiarism or academic integrity breach does not depend on students' use of AI tools but depends on whether or not students have disclosed this utilization. In relation to this, many scientific journals now require that the authors declare generative AI and AI-assisted technologies. In my alma mater, Princeton University, we had no proctors in examinations. We wrote and signed "I pledge my honor that I have not violated the honor code during this examination" in all exams, and during my 4 years of undergraduate education there, I did not encounter a single incident of cheating. As this culture of integrating not cheating into one's honor seems to have worked for centuries in Princeton University, it can also be possible to integrate using AI in the designated and appropriate ways for assignments into one's honor and moral understanding of themselves.

To achieve this purpose, I would recommend that teaching institutions create a constitution of rules regarding the appropriate conditions for using AI, to which the students can then pledge allegiance by signing the honor code. When talking about signing and honesty, the once highly influential Shu et al. (2012) study comes to mind, though it only mentions signing one's name with no reference to an honor code. Their findings stated that signing one's name at the beginning as opposed to the end of a document decreased dishonesty (Shu et al., 2012), but in light of evidence that questioned the validity of their data, this article was retracted in 2021.[1] Yet, just because the validity of this study's data is questioned does not mean that signing an honor code will not have an effect. Nonetheless, the most effective ways to induce ethical behavior regarding AI usage could be scientifically investigated in today's educational settings, within the context of the current ethical values and understanding, so that the ideal ways to proceed can be revealed. This seems to be a promising avenue, as an analysis of the last three decades of research on honor codes concludes that honor codes are effective in "creating and maintaining a culture of academic honesty and integrity in institutions of higher education" (Tatum, 2022).

Institutions are beginning to address the ethical use of AI in academic contexts, though a formal honor code on this does not yet exist. For instance, in their "Syllabus and Honor Code Statements Regarding the Use of AI," Saint Leo University declares that "it is important to use AI appropriately and ethically,"[2] but does not specify how to do so other than showing how to cite AI sources. They add to their Academic Misconduct statement a clause of "Unapproved use of Artificial Intelligence," the only explanation of which is "Unauthorized student use of AI-powered tools or services to enhance performance on exams, tests, quizzes, discussion posts, papers, projects, or other assignments"[2] with no specifications or details. Similarly, the University of Colorado Boulder states unapproved use of AI to be considered cheating and to violate the honor code and warns the students that using AI outside of the standards of the course will be considered as violating the honor code.[3] Yet, they do not give a clear guideline on what constitutes such violations and say that "CU Boulder does not have a policy on AI in curriculum because there is no one-size-fits-all approach for using AI tools on our campus" and that students are "responsible for working with [their] professors to understand their guidelines and clarify questions before using AI for coursework."[3] Students being encouraged to learn AI usage expectations specific to each course from their professors indicates no uniformity and leaves the definition of ethical AI usage to the discretion of each professor. Such disparities across courses may be confusing to students, where what is ethical for one professor may not be for another, precluding the students from forming a clear and consistent understanding of what it means to use AI ethically at school. Though warnings made by universities to use AI ethically and in line with the honor code are important, they are still vague. This unclarity underlines the need for a standardized and clear honor code to be generated for the ethical use of AI in academic contexts.

There are certain guidelines to implement the honor code specific to the ethical utilization of AI that I propose here. McCabe, Trevino, and Butterfield (2002) have shown that modified honor codes were associated with lower levels of cheating behavior. While they note that the modified honor code does not have a single definition, they highlight that it includes two strategies: the institution clearly communicating to students that academic integrity is a priority, and students being given a significant role in the judicial body and in the propagation of the purposes and enforcement of the code (McCabe et al., 2002). These strategies can be applied to an AI-focused honor code, where students can be given initiative in helping develop the code along with the institution, following the suggestions of McCabe et al. (2002), alongside the responsibility of the institution in clearly communicating that ethical use of AI is an institutional priority.

In addition to formulating a standardized honor code for AI usage, educational institutions can adapt to the AI revolution by teaching students how to use AI tools in line with the cognitive objectives of higher education. What I mean here is not simply that curricula should include teaching of how to use AI tools, but that educators should teach students to use AI in a way that is conducive for their analytical and critical thinking, instead of a way that hampers and takes these skills away from them. For instance, educators can make students understand that they can use AI to find the most suitable sources for their topic but should utilize their own intelligence and ability to synthesize content to bring these materials together. These teachers can conduct practice exercises of researching with AI and synthesizing on their own to answer research questions and assignments. By such means, using AI in a way that does not interfere with human intelligence and its development may be possible. A similar observation has been made by Harvard Graduate School of Education lecturer on education, Houman Harouni. He believes that educators must guide students on how to use AI tools responsibly and "develop instruments and ways of navigating this reality with integrity" (Ross, 2023), a sentiment that parallels the suggestions I make here. Specifically, he wants educators to train students in asking questions and to teach students how to apply critical thinking to their own questions and to the answers provided by AI (Ross, 2023). Integrating AI into education is, of course, not a novel idea, and this is why I am not simply recommending that we teach the use of AI technologies in schools. My proposal in this paper is that we teach students how to use AI ethically and in a way that does not hinder but fosters creative and analytical thinking.

The shift of focus I suggest for educational institutions from better detecting AI-generated text in student submissions to educating students on how to use AI in ways that are conducive to higher-level thinking and learning parallels Cowen's (2009) interpretation of the Turing Test, i.e. "Turing cleverly shows why the supposed test is misleading and the real question is how to educate both children and machines, not how to distinguish them," and marks how Turing notes that "much needs to be done, presumably in both computer education and in human education" (Cowen and Dawson, 2009). Of course, machine learning and teaching AI to be more helpful to humans is very

important, yet just as important is the mission to educate children about the ideal ways of AI utilization in this fast-evolving technological landscape in which they ought to survive and thrive.

## Conclusion

The ability of new natural language generation algorithms to produce fluent, grammatical texts has raised important concerns about how they will relate to human cognition as well as the ethics behind their utilization. In this article, I explored the intersection of human cognition and AI language generators through how humans think about AI and how they evaluate their thoughts about AI (metacognition). I then proposed the formation of an honor code specific to the ethical utilization of AI in order for AI language generators to not take away the cognitive gains that the educational systems are designed to provide to students and instead, facilitate and support them.

Given that one characteristic of metacognitive evaluation is overconfidence in one's ability to detect artificially generated content (Köbis et al., 2021), we may actively try to be more aware of our limitations in detecting AI-generated texts. Additionally, by being aware of the limitations of AI-generated texts with regard to accuracy and depth, we can work towards improving AI with our feedback so that AI can serve us even better in the future. This perspective is a more productive one compared to the fear that AI will take over all professions and human cognition, where, as stated by Chistopher Losee (February, 2023), the more likely future scenario is that "AI will not replace you, people using AI will."

While applying this collaborative approach to educational settings, inculcating a culture of ethical AI utilization will be paramount, as creating this culture of honesty has implications for the prosperity of societies. This is because by using AI ethically and not having a language-generating system do their work for them, the future leaders, decision-makers, and professionals of specialized occupations will actually learn their studied materials and excel in their positions to serve society to the best of their ability.

As Jakesch et al. (2023) highlight, we need to ensure that AI language systems "support rather than undermine human cognition." Machine learning is an exciting avenue in many domains, but focusing on improving how machines learn should not preclude humans from learning, who should be educated in appropriate and ethical ways of utilizing machines for a prosperous future.

**Declaration of generative AI and AI-assisted technologies in the writing process**. During the preparation of this work, no generative AI or AI-assisted technologies were used for the writing and editing of this article. Upon acceptance of the article, QuillBot was used to carry out the editorial request for a final language check.

## Data availability

This manuscript has no associated data.

Received: 18 June 2024; Accepted: 24 October 2024;
Published online: 18 December 2024

## Notes

1 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8463784/
2 https://faculty.saintleo.edu/wp-content/uploads/2023/10/AI-Syllabus-and-Honor-Code-Statements-Aug-2023.pdf
3 https://www.colorado.edu/studentaffairs/2024/09/16/ai-and-honor-code-3-things-know

## References

Cowen T (2009) What does the Turing test really mean? Marginal Revolution, available at https://marginalrevolution.com/marginalrevolution/2009/07/what-does-the-turing-test-really-mean.html
Cowen T, Dawson M (2009) What does the Turing test really mean? And how many human beings (including Turing) could pass? George Mason University, available at https://d101vc9winf8ln.cloudfront.net/documents/28495/original/turingfinal.pdf?1529586155
Fink J (May, 2024) Bluesky post retrieved on September 6, 2024 from https://bsky.app/profile/planetoffinks.bsky.social/post/3kshn22j6hm2e
Frith CD (2012) The role of metacognition in human social interactions. Philos Trans R Soc B: Biol Sci 367(1599):2213–2223
Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for AI-generated language are flawed. Proc Natl Acad Sci USA 120(11):e2208839120
Kahneman D (2013) Thinking, fast and slow. Farrar, Straus and Giroux, New York, NY
Köbis NC, Doležalová B, Soraperra I (2021) Fooled twice: people cannot detect deepfakes but think they can. Iscience 24(11):103364
Köbis N, Mossink LD (2021) Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. Comput Hum Behav 114:106553
Kralik JD, Lee JH, Rosenbloom PS, Jackson Jr PC, Epstein SL, Romero OJ, McGreggor K (2018) Metacognition for a common model of cognition. Procedia Comput Sci 145:730–739
Krishna K, Song Y, Karpinska M, Wieting J, Iyyer M (2024) Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. Adv Neural Inf Process Syst 37:1–32
Losee C (February, 2023) AI will not replace you, people using AI will. Retrieved on April 11, 2023 from https://www.linkedin.com/pulse/ai-replace-you-people-using-christopher-losee/?trk=pulse-article_more-articles_related-content-card
Ma Y, Liu J, Yi F (2023) Is this abstract generated by AI? a research for the gap between AI-generated scientific text and human-written scientific text. https://arxiv.org/abs/2301.10416
McCabe DL, Trevino LK, Butterfield KD (2002) Honor codes and other contextual influences on academic integrity: a replication and extension to modified honor code settings. Res High Educ 43:357–378
Mireshghallah N, Mattern J, Gao S, Shokri R, Berg-Kirkpatrick T (2024) Smaller language models are better zero-shot machine-generated text detectors. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. Vol. 2, Short Papers. Association for Computational Linguistics, pp 278–293
Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C (2023, July) Detectgpt: Zero-shot machine-generated text detection using probability curvature. In: International Conference on Machine Learning. PMLR. pp. 24950–24962
Mollick E (April, 2023) LinkedIn post retrieved on April 11, 2023 from https://www.linkedin.com/posts/emollick_teachers-should-be-wary-of-ai-detection-tools-activity-7047219085985660929-K99C?utm_source=share&utm_medium=member_desktop
Perkins M (2023) Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. J Univ Teach Learn Pract 20(2):07
Ross EM (2023) Embracing artificial intelligence in the classroom: generative AI tools can reflect our failure of imagination and that is when the real learning starts. Harvard Graduate School of Education. Retrieved on Oct 15, 2024 from https://www.gse.harvard.edu/ideas/usable-knowledge/23/07/embracing-artificial-intelligence-classroom
Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S (2023) Can AI-generated text be reliably detected? https://arxiv.org/abs/2303.11156
Shu LL, Mazar N, Gino F, Ariely D, Bazerman MH (2012) Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. Proc Natl Acad Sci USA 109(38):15197
Tang R, Chuang YN, Hu X (2024) The science of detecting llm-generated text. Commun ACM 67(4):50–59
Tatum HE (2022) Honor codes and academic integrity: Three decades of research. J Coll Character 23(1):32–47
Turing AM (1950) Mind. Mind 59(236):433–460
Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. Science 185(4157):1124–1131
Watt S (1996) Naive psychology and the inverted Turing test. Psycoloquy 7(14):463–518

## Acknowledgements

I sincerely thank Mr. Philip Gee for his diligent proofreading and invaluable suggestions that significantly enhanced the quality of this paper. I am grateful to Mr. Mert Pamukcu

for running my accepted article through QuillBot before publication to check for grammatical errors.

## Author contributions

The author CD is responsible for the conception of this comment article, for writing the initial submitted version, for implementing the revisions, and for preparing the final version of the manuscript for publication.

## Competing interests

The author declares no competing interests.

## Ethical approval

Ethical approval was not required as the article does not involve human participants.

## Informed consent

This article does not contain any studies with human participants performed by any of the authors.

## Additional information

**Correspondence** and requests for materials should be addressed to Ceymi Doenyas.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024