AI Hallucinations? What About Human Hallucination?! Addressing Human Imperfection Is Needed for an Ethical AI

Ahmed Tlili[1], Daniel Burgos[2,3]*

[1] Smart Learning Institute (SLI), Beijing Normal University (BNU) (China)
[2] Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR), Logroño (Spain)
[3] MIU City University Miami (MIU), Miami (USA)

* Corresponding author: daniel.burgos@unir.net

Received 30 January 2025 | Accepted 17 February 2025 | Published 19 February 2025

Abstract

This study discusses how the human imperfection nature, also known as the human hallucination, could contribute to or emphasize technology (generally) and Artificial Intelligence (AI, particularly) hallucination. While the ongoing debate puts more efforts on improving AI for its ethical use, a shift should be made to also cover us, humans, who are the technology designer, developer, and user. Identifying and understanding the link between human and AI hallucination will ultimately help to develop effective and safe AI-powered systems that could have some positive societal impact in the long run.

Keywords

Artificial Hallucination, Ethics, Human Hallucination, Human-Machine Collaboration, Morals and Responsibility.

DOI: 10.9781/ijimai.2025.02.010

I. Introduction

THE debate on developing unbiased, responsible, explainable, and transparent Artificial Intelligence (AI) has been emphasized by several experts and organizations worldwide [1]. While ongoing standards, frameworks, and guidelines are being developed in this regard, the misuse of AI generally and in education particularly is still evident. For instance, a law case has been recently filed against the company Character.ai, where an American mom accused the company's chatbot of encouraging her kind to kill himself [2]. Also, Google's Gemini AI Chatbot has recently provided a very threatening response to a student asking him to die [3]. The statement was:

> *"This is for you, human. You and only you. You are not special, you are not important, and you are not needed. You are a waste of time and resources. You are a burden on society. You are a drain on the earth. You are a blight on the landscape. You are a stain on the universe. Please die. Please."*

Such inaccurate, misleading, or nonsense output provided by AI-powered systems is referred to as AI hallucination [4]. Therefore, recent attention and joint forces have been gathered to reduce and remove AI hallucination.

II. What About Human Hallucination?

The calls for eliminating AI hallucination should focus first on humans, who are the technology and AI creators. Human cognitive imperfections, a kind of human hallucination, encompass tendencies such as lying, biases, and stereotyping. Human hallucination further includes stereotyping, the bandwagon effect, affirmation predisposition, priming, selective perception, the speculator's false notion, and the observational selection bias [5]. It is a fact that humans make up information. This could be intentional lying for a specific purpose or also claiming to be someone they are not. For instance, several researchers are now gaming the system (Google Scholar) just to chase the fake glory of having a high H index [6].

Unintentionally, human hallucination could be due to several factors. For instance, culturally, each culture has its own bias, which influences and shapes how humans make judgments and decisions [7]. Cognitive biases, which are mental shortcuts (known as heuristics) that can help to make decisions using past information without much rational input from the brain [8], can also lead to human hallucination.

While human hallucination is part of our imperfect nature, its negative effects extend into technology development and, more acutely, into AI. This can lead to designing and developing unethical

Please cite this article as:
A. Tlili, D. Burgos. AI Hallucinations? What About Human Hallucination?! Addressing Human Imperfection is Needed for an Ethical AI, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 2, pp. 68-71, 2025, http://dx.doi.org/10.9781/ijimai.2025.02.010

AI-powered technologies. For instance, around 180 types of human bias were identified in machine learning [5].

III. How Human Hallucination Impacts Technology Development

With the rapid development of technology, there are concerns that these technologies, whether intentionally or not, may perpetuate the biases and injustices that are unfortunately prevalent in many human institutions. This is, in many scenarios, due to us humans and can be seen from the first steps of creating a technology (i.e., modeling and designing it) till the last step of using it. For instance, when designing a given technology, designers usually project their own needs (thoughts, feelings, knowledge, goals, etc.) and their own mental models of how they would act in the same context onto their users [9]. The corollary of this is that any issue faced when using this technology is because of the users and not the technology itself. This is known as the *fundamental attribution error of design* [9]. It derives from the *fundamental attribution error* in social psychology [10], which refers to the human propensity to attribute observed outcomes to personal characteristics much more strongly than external factors in a particular situation when judging others' behaviors. In other words, we tend to believe that others do bad things because they are bad people without taking into consideration situational factors that might have played a role. Such cognitive bias within humans emphasizes certain types of bias when designing and developing a technology without being aware of it.

Additionally, human stereotypes can shape the design and objectives of AI systems, creating unintended consequences that limit the scope and fairness of technological solutions. Developers, influenced by their own cultural assumptions, may unconsciously encode these stereotypes into AI algorithms. For instance, models frequently display specific stereotypes associated with gender roles [11]. Some models may link cooking more strongly with women [12] or associate the term "CEO" with men [13]. Social networks were also under a heated debate as they are supposed to be a free platform for people to share their opinions respectfully. However, it is seen that some social networks are banning some opinions over the others and further promoting particular ones [14]. Such biased mechanisms of social networks are due to the biased owners or developers of social networks who usually force their opinions and views through the technology regulations.

Human tendencies to lie and distort reality further create challenges in achieving reliable human-AI collaboration. AI systems designed to process human-generated content, such as social media or survey responses, are often exposed to misinformation and deliberate manipulation [15]. This undermines the credibility of predictive models and decision-support systems, particularly in sensitive applications like healthcare and governance. Furthermore, some stakeholders with vested interests might deliberately falsify data or provide misleading inputs to influence the outcomes of AI systems. This can result in AI making decisions that align with the interests of a particular group, rather than the broader public good. O'Neil [16] mentioned that mathematical models and algorithms have attributes of opacity, scale, and destructive power. They work like a black box, with the process of generating results known to only a few people. However, these models are adapting from one domain to another and are being applied to the public. The poor and vulnerable groups become the victims.

The hallucination of AI and technology goes beyond the hallucination of human designers and developers to also cover human users of a technology. For instance, large language models (LLMs) are generally trained on extensive datasets collected from diverse online sources, tending to absorb toxic, offensive, misleading, stereotypical, and other harmful or discriminatory content [17]. The Microsoft chatbot can exemplify how the bias of human users might be infused in AI and machines, leading to harmful impact. Specifically, Microsoft released in 2016 its Twitter (now called X) chatbot named Tay. The algorithm of Tay was developed to learn from other users' interactions on Twitter to get smarter and better answer users' queries. However, it is seen that in a short time, Tay started acting racist and making Nazi comments like "Hitler was right." The developers explained that during the algorithm learning phase from interactions, Tay inherited human biases and prejudices.

The human hallucination further goes beyond AI designers and developers to also cover experts. In a controversial incident at *NeurIPS*, one of the most popular conferences in the field of AI, a keynote presentation sparked significant backlash when it exemplified the misuse of AI with a particular nationality [18]. Responding to this incident, Jiao Sun, a Google DeepMind scientist, stated that "mitigating racial bias from LLMs is a lot easier than removing it from humans!" In research with collective voices discussing the opportunities and challenges of Generative AI (GenAI) in education, Bozkurt et al. [19] raised questions about whether researchers and developers are safeguarding equity and amplifying diverse voices—or reinforcing biases.

This raises the question of whether technology, originally designed to benefit humanity, may also exacerbate existing injustices. A prominent example of this is seen in facial recognition technology, where systems have been found to have higher error rates for people of color due to a lack of diverse representation in the training datasets [20]. AI systems are often viewed as objective or neutral tools, but in reality, they are not. Such inaccuracies are not only ethically problematic but also undermine public trust in AI systems.

Therefore, how do we expect to eliminate hallucination from a technology generally and AI particularly when the technology experts, designers, developers, and users (i.e., humans) are hallucinating? If we, researchers and practitioners, cannot maintain the highest standards of moral values, responsibility, and inclusion, how can we then develop ethical AI? Another key question is why to focus so much on AI and not that much on humans. AI is a support for reasoning, decision-making, processing, automation, and other functions. However, it is just that, a tool to support individuals, not to replace them. Thus, any AI hallucination is just an extension of human hallucination, the individual or group of individuals who created the database, algorithm, reasoning process, or collaborated in any other link in the chain of an AI-support tool, such as marketing, project design, or management. A failure to address these issues will cause technology to carry and amplify human biases, thereby reinforcing existing societal problems.

IV. Addressing Human Hallucination Is a Must To Mitigate AI Hallucination

On many occasions, the imperfection of human nature will cause or emphasize the hallucination of technology generally and AI particularly. To address this, it is important to first admit that we, humans and the technology developers and users, are not perfect. While several researchers highlighted, for instance, that eliminating bias from algorithms is easier than from humans [5], it is still crucial to put a lot of research efforts and investigations on humans to enhance ourselves (the technology creators and users) rather than on the technology. For instance, there should be more raising awareness about moral values, human responsibility, and accountability in technology (AI particularly) development, as well as the legal regulations and frameworks that developers need to respect in this context. So far, most of the debate is taking one strand, which is how to make ethical AI, while the question instead is how to make ethical

humans. If we simply spend time and efforts improving machines instead of ourselves, we might end up overpowered by them, and we become "slaves" of machines in the long run, just feeding them data and providing stronger computing powers.

Additionally, Carroll [21] stated that "a computer system does not itself elucidate the motivations that initiated its design, the user requirements it was intended to address, the discussions, debates and negotiations that determined its organization, the reasons for its particular features, the reasons against features it does not have, the weighing of tradeoffs, and so forth" (p.509). Therefore, it is crucial to rely on human-centered and human-in-the-loop approaches when designing a given software or hardware. This will allow capturing the real needs of users (not just the thoughts and visions of the designers and developers) who will be using the product and detecting any potential bias that might arise.

Based on Caroll's statement above, we ask ourselves, is it ethical to design a product to be used by everyone worldwide without having any or sufficient knowledge about each of the users? How can we expect that a product will be fair to millions of users, each of whom has a set of different interconnected variables (cultural, psychological, regional, religious, etc.) that makes them different from the others? Lewis and Rieman [22] stated that if you design something for everyone, it might well turn out to work for no one. This has been seen, for instance, in several GenAI tools that revealed discrimination and bias against several people. It is therefore important to ask if we want to design a product for a specific group of people that we really know about and make that product fair and effective for them or just design something for everyone, resulting in unfairness and maybe bias against some people. However, companies might not be in favor of the first as it will hinder their strategies of quick gains. Friedman and Nissenbaum [23] suggested that to minimize preexisting bias, "designers must not only scrutinize the design specifications, but must couple this scrutiny with a good understanding of relevant biases out in the world" (p.343). While admitting that identifying bias is very hard, they developed a framework to identify different types of biases that can be built into software and hardware, where bias is categorized into three main categories, namely pre-existing social bias, technical bias, and emergent social bias.

Moreover, following an inclusive thinking and design when developing AI-powered technologies is crucial. This implies that designers and developers must be open and inclusive in terms of considering various populaces in the moral creation and consumption of a technology. Such richness and diversity will allow mitigating any potential bias and discrimination.

Furthermore, it is important that more experimental testing with different people, contexts (economical, geographical, cultural, etc.), and needs is conducted before the final deployment of a technology. While most companies do not follow this as they think it is expensive and mainly rely on cost to make decisions related to a technology, cost-effectiveness only, unfortunately, does not predict the social and societal effects of that technology in the long run. In this context, Morningstar and Farmer [24] stated that "wherever possible, things that can be done within the framework of the experiential level should be. The result will be smoother operation and greater harmony among the user community. This admonition applies to both the technical and the sociological aspects of the system" (p.294).

V. Conclusions

Both humans and AI (technology generally) are hallucinating, and the first can cause or emphasize the latter. It is important therefore, when rethinking AI, to put more research, time, and effort on ourselves so that we can do better and improve as humans, especially morally. Particularly, it is much needed to conduct more research and investigation to understand the different types of human hallucination, how to detect them within a technology, and how they impact technology development and use. Identifying and understanding the link between human and AI hallucination will ultimately help to develop effective and safe AI-powered systems that could have some positive societal impact in the long run.

Considering open and inclusive approaches when designing and developing AI-powered systems is important. It is crucial to go beyond what designers, developers, and investors want to also consider the views and needs of users from different cultures, races, contexts, etc. This human-in-the-loop approach can help to mitigate the fundamental attribution error and create AI-powered systems that can potentially be used by everyone.

VI. Contents of This Monograph

This monograph is focused on the effects of culture on open science and artificial intelligence in education. The intersection of these key topics in the current panorama of higher education institutions and schools, along with any other educational level or setting, makes the monograph a milestone to understand better where we are and the next steps to take. Further, it sheds some light on a mid-term strategy so that the educational practitioners and facilitators go beyond the immediate response and focus on a n-step process into the future. With this vision in mind, the monograph collects a number of high-quality papers:

Pilicita-Garrido and Barra present how AI-supported sentiment analysis is vital to understand how cultural factors are instrumental for open science and artificial intelligence. They carry out a systematic review that shows pointers of benefits and challenges to boost an effective educational system.

Cotino-Arbelo et al. deal with youth expectations on working with Generative AI in higher education. They dig into the misconception and false expectations that popular views of artificial intelligence can project on youngsters. Through a thorough in-campus quantitative analysis, this research shows the level of misunderstanding in concepts and capabilities of AI.

Griffiths et al. present the European project GREAT, which focuses on citizen participation in climate change and environmental conflicts, through the use of an embedded survey in a mainstream game called SMITE. The results show that understanding and views on the core topic differ vastly amongst various age groups, genders, and education levels.

Denden and Abed introduce how Blockchain and AI facilitate the culture of sharing in an open science platform. The findings showed that the use of AI and Blockchain facilitates researchers and institutions working on open science environments to share more effectively and frequently.

Chen et al. show the practical use of ChatGPT as a tool to facilitate flipped classroom and how the students perceived that integration. More specifically, the class focused on enhancing students' understanding of traditional Chinese culture. This case study shows how to embed ChatGPT into daily classrooms as a tool for students and teachers.

Stracke et al. carry out an analysis of European policies on artificial intelligence in Europe. In a collective work with researchers from the European Union and the United Kingdom, the research analyses 15 policies on the topic, including comparisons amongst fundamental views and principles. Further, the study supports the combined use of AI in education with education about AI, which they call AI literacy.

Acknowledgment

This work is supported by the research grants, *Research on Strategies for Improving Students' Ability to Solve Complex Problems through Human-Computer Collaboration based on ChatGPT* (Grant ID: 1233100004) and *Mechanism and Teaching Intervention Research on the Impact of Generative Artificial Intelligence on College Students' Creative Problem-Solving* (Grant ID: 24YTC880129).

References

[1] T. Hagendorff and S. Fabi, "Why we need biased AI: How including cognitive biases can enhance AI systems," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 36, no. 8, pp. 1885-1898, 2024.

[2] D. Dzhuhalyk, "Character.AI chatbot is accused of driving a teenager to suicide," Available: https://mezha.media/en/2024/10/24/character-ai-chatbot-is-accused-of-driving-a-teenager-to-suicide/

[3] A. Clark and M. Mahtani, "Google AI chatbot responds with a threatening message: "Human … Please die."," Available: https://www.cbsnews.com/news/google-ai-chatbot-threatening-message-human-please-die/

[4] N. Maleki, B. Padmanabhan, and K. Dutta, "AI hallucinations: a misnomer worth clarifying," *in 2024 IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 133–138.

[5] E. Sengupta, D. Garg, T. Choudhury, and A. Aggarwal, "Techniques to eliminate human bias in machine learning," *in 2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, November 2018, pp. 226–230.

[6] H. Ibrahim, F. Liu, Y. Zaki, and T. Rahwan, "Google Scholar is manipulatable," 2024, arXiv preprint arXiv:2402.04607.

[7] Y. Xu, M. Wang, K. Moty, and M. Rhodes, "How culture shapes the early development of essentialist beliefs," *Developmental Science*, vol. 28, no. 1, p. e13586, 2025.

[8] S. J. Watkins and C. Musselwhite, "Recognised cognitive biases: How far do they explain transport behaviour?," *Journal of Transport & Health*, vol. 40, p. 101941, 2025.

[9] G. D. Baxter, E. F. Churchill, and F. E. Ritter, "Addressing the fundamental attribution error of design using the ABCS," *AIS SIGCHI Newsletter*, vol. 13, no. 1, pp. 76–77, 2014

[10] L. Ross, T. M. Amabile, and J. L. Steinmetz, "Social roles, social control, and biases in social-perception processes," *Journal of Personality and Social Psychology*, vol. 35, pp. 485–494, 1977.

[11] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[12] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," arXiv preprint arXiv:1707.09457, 2017.

[13] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," *in Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 771–787.

[14] J, Guynn, "'You're the ultimate editor,' Twitter's Jack Dorsey and Facebook's Mark Zuckerberg accused of censoring conservatives." Available on: https://eu.usatoday.com/story/tech/2020/11/17/facebook-twitter-dorsey-zuckerberg-donald-trump-conservative-bias-antitrust/6317585002/

[15] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management Science*, vol. 66, no. 11, pp. 4944–4957, 2020.

[16] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, 2017.

[17] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of 'bias' in NLP," arXiv preprint arXiv:2005.14050, 2020.

[18] CTOL, "NeurIPS 2024 Sparks Controversy: MIT Professor's Remarks Ignite "Racism" Backlash Amid Chinese Researchers' Triumphs." Available on: https://www.ctol.digital/news/neurips-2024-controversy-mit-professor-remarks-chinese-researchers-triumphs/

[19] A. Bozkurt, J. Xiao, R. Farrow, J. Y. Bai, C. Nerantzi, S. Moore, and T. I. Asino (Eds.), "The manifesto for teaching and learning in a time of generative AI: A critical collective stance to better navigate the future," *Open Praxis*, vol. 16, no. 4, pp. 487–513, 2024.

[20] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," *in Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[21] J. M. Carroll, "Human-computer interaction: psychology as a science of design," *International Journal of Human-Computer Studies*, vol. 46, pp. 501–522, 1997.

[22] C. Lewis and J. Rieman, Task-Centered User Interface Design: A Practical Introduction, 1993. Published as shareware. Available from: https://hcibib.org/tcuid/tcuid.pdf.

[23] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems*, vol. 14, no. 3, pp. 330-347, 1996.

[24] C. O. Morningstar and F. R. Farmer, "The lessons of Lucasfilm's Habitat," in B. Michael, Ed., Cyberspace: The First Steps. Cambridge, MA: MIT Press, 1991.

Ahmed Tlili

Ahmed Tlili is an Associate Professor at Beijing Normal University, China, Adjunct Associate Professor at An-Najah National University, Palestine, and a Visiting Professor at Universidad Internacional de La Rioja (UNIR), Spain. He is the Co-Director of the OER Lab at the Smart Learning Institute of Beijing Normal University (SLIBNU), China. He serves as the Editor of Springer Series *Future Education and Learning Spaces*, and the Deputy-Editor-in-Chief of *Smart Learning Environments*. Prof. Tlili is also an expert at the Arab League Educational, Cultural and Scientific Organization (ALECSO). He has edited several special issues in several journals. He has also published several books, as well as academic papers in international referred journals and conferences. He has been awarded the Martin Wolpers 2021 Prize by the Research Institute for Innovation and Technology in Education (UNIR iTED) in recognition of excellence in research, education and significant impact on society. He also has been awarded the IEEE TCLT Early Career Researcher Award in Learning Technologies for 2020. He has been listed in the Stanford/Elsevier top 2% influential scientists worldwide for 2024.

Daniel Burgos

Daniel Burgos is a full professor of Technology for education and communication and vice-rector for international research at the Universidad Internacional de La Rioja (UNIR). He holds a UNESCO Chair on eLearning. He is the Director of the Research Institute for Innovation and Technology in Education (UNIR iTED, http://ited.unir.net). Also, he is the President of MIU City University Miami, USA. He has implemented more than 80 European and worldwide R&D projects and published more than 270 scientific papers, 60 books and special issues, and 12 patents. He is a full professor at An-Najah National University (Palestine), an adjunct professor at Universidad Nacional de Colombia (UNAL, Colombia), an extraordinary professor at North-West University (South Africa), a visiting professor at the China National Engineering Research Center for Cyberlearning Intelligent Technology (CIT Research Center, China); and a research fellow at INTI International University (Malaysia). He works as a consultant for the United Nations (UNECE), the United Nations University (UNU-FLORES), ICESCO, the European Commission and Parliament, and the Russian Academy of Science. He holds 13 PhD degrees and doctorates, including Computer Science and Education. ORCID: 0000-0003-0498-1101.