HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240194

Evaluating Generative AI Incidents: An Exploratory Vignette Study on the Role of Trust, Attitude and AI Literacy

Esther S. KOX[a,b,1] and Beatrice BERETTA[a]
[a] *University of Twente, Enschede, The Netherlands*
[b] *TNO, Soesterberg, The Netherlands*
ORCiD ID: Esther Kox https://orcid.org/0000-0002-2512-5665

Abstract. Generative AI presents vast opportunities but also risks. Misuse, whether intentional or not, can lead to significant “real-world” consequences. We presented subjects (n=139) with five vignettes describing incidents involving generative AI. We explored the relationship between their level of AI literacy, attitude towards AI, trust in AI chatbots, and people’s reactions to the vignettes. Attitude and trust, measured before and after the vignettes, declined significantly. However, these changes as well as the reactions to the vignettes were unrelated to AI literacy. Yet, higher AI literacy was associated with more frequent use of AI chatbots, higher trust and more positive attitudes towards AI. So while AI literacy appeared to be related to the general perceptions and usage of generative AI, it was not linked to the evaluation of incidents involving generative AI. The implications for trust calibration and appropriate reliance are discussed.

Keywords. LLMs; AI chatbots; trust; attitude; AI literacy

1. Introduction

Artificial Intelligence (AI) technology has revolutionized the way we handle information. Especially generative AI models (e.g., ChatGPT) are seen as one of the most disruptive technological breakthroughs in recent years [1,2]. Since its release end of 2022, ChatGPT, an AI chatbot developed by OpenAI, has gradually gained popularity and has changed how people perceive and interact with AI [3,4]. Models like ChatGPT are Large Language Models (LLMs); a type of AI designed to understand and generate language in a way that mimics human language processing abilities [5]. These models are trained using immense amounts of publicly available text from the internet, exposing the model to a wide range of topics, writing styles, and linguistic patterns [5], which enables them to capture the nuances of human language and to produce highly coherent and human-like responses [2,6,7]. These qualities make them perfect for conversational use and have contributed greatly to their popularity. Generative AI offers many potential and realized benefits for people, organizations and society in a wide range of sectors [2,8].

[1] Esther S. Kox, esther.kox@tno.nl

However, the use of generative AI also carries several risks [2,6]. First, generative AI can produce harmful and inappropriate content (e.g., discriminative content, promotion of harmful ideologies) [2]. Both the model's training data and its lack of contextual understanding can lead it to generate content that is considered inappropriate or culturally insensitive. Second, generative AI can fabricate fictitious or erroneous content with a high level of plausibility [9]. In literature, this phenomenon is often referred to as “hallucination” or “confabulation” [10,11], but we will use the non-anthropomorphic term *fabrication* [2,12]. Third, the increasing difficulty in determining the authenticity of information and media (e.g., Deepfake) [2]. Fourth, training data for algorithms often contain biases, unfairly favouring or disadvantaging certain individuals or groups, which can seep into the model's output, sustaining societal biases [1,2].

These issues can cause a wide range of real-life consequences [6,13]. For example, biased algorithms can, when deployed and acted upon, exacerbate existing societal inequality. Also, if patients, for example, rely on fabricated information for medical advice, it could result in life-threatening situations [5,6,14]. Alternatively, the manipulation of audio-visual media can exacerbate the far-reaching effects of misinformation on social media., such as polarisation [15]. The possible instances of intentional and unintentional misuse are numerous [5]. To mitigate the risks and maximize the benefits associated with generative AI, people must understand how AI works and is applied, so that they can trust in and rely on it appropriately.

1.1. Trust

During interaction, people continuously adjust their level of trust in an artificial agent based on their ongoing interactions and experiences with the aim to align the perceived trustworthiness of an agent with their actual trustworthiness (i.e., trust calibration) [16]. Consistent with that objective, people are increasingly (made) aware of the fact that AI chatbots can provide misleading information that could harm user’s interests or well-being (e.g., false medical, legal, or financial advice) [6]. For instance, OpenAI added a warning to ChatGPT’s main screen: “ChatGPT can make mistakes. Verify important information.”. Nevertheless people seem to trust in and rely on AI more than they should.

Prior work shows how people are often misled by incorrect AI predictions and how they would, in some cases, make better decisions on their own [17,18]. Trusting false AI generated output can have major societal implications. For instance, a lawyer depended on ChatGPT to draft a motion replete with fabricated case law, because he “did not comprehend that ChatGPT could fabricate cases” [19] (see vignette ‘Fabrication’, Table 1). People often tend to follow the advice of automation, without verifying it, because they consider a machine as infallible (i.e., automation bias) [7,20]. Adding a warning that ChatGPT can make mistakes is an attempt to prevent overreliance. However, the effect of the warning is likely to be nullified by the convincing and seemingly sophisticated output that ChatGPT generates.

ChatGPT’s ability to generate highly coherent answers “can fool us into thinking that they understand more than they do” [21]. ChatGPT’s output seems highly plausible and intelligent, but lacks comprehension [22]. However, the human-like way of communicating triggers people’s tendency to attribute humanlike capabilities to non-human entities (i.e., anthropomorphism) [23]. People tend to base their level of trust on attributed characteristics rather than on actual experiences with the agent itself [24], creating a discrepancy between the perception and its actual capabilities [6]. As such, anthropomorphism can lead to misplaced trust and inappropriate reliance.

Some scholars therefor argue that anthropomorphic features in the design of artificial agents should be avoided [17,24,25]. However, despite the potential risks, the human-like responses are exactly what contributed to the success and ease-of-use of AI chatbots. But not all people are equally susceptible to anthropomorphic cues [23,26]. Researchers have proposed that people who lack AI literacy tend to anthropomorphize AI-agents more [6]. People with a limited understanding of AI may more easily fall for the illusion of intelligence and overestimate it based on superficial interactions. Increasing people's AI literacy could mitigate the risks associated with anthropomorphic design, while holding on to its benefits.

1.2. AI literacy

AI literacy is defined as a broad set of skills that enable individuals to recognize everyday applications of AI, know the basic functions of AI and understand how to use AI effectively in daily life [27]. For many, AI is still a "black box" with difficult to determine opportunities and risks [28]. AI literacy enables individuals to make informed decisions about AI [29] and help people to gauge when it is appropriate to rely on it [8]. For example, research shows that clinicians with higher AI literacy were less likely to rely on incorrect medical AI recommendations than clinicians with lower AI literacy [18]. AI literacy is thought to counter biases that are known to interfere with accurate trust calibration, appropriate reliance and effective decision-making (e.g., anthropomorphism, automation bias). As such, understanding AI's strengths and weaknesses is deemed crucial for mitigating instances of misuse and deception like mentioned earlier [6,16,30]. More research is needed to determine how AI literacy affects people's trust in, attitude towards and perceptions of AI across different contexts.

1.3. Current study

The aim of this study was to explore how people respond to scenario's describing incidents involving generative AI and whether these responses are influenced by an individual's level of AI literacy. It was expected that trust in and attitudes towards AI of people with higher level of AI literacy would be less affected by the vignettes in comparison to people with lower AI literacy. This stems from the idea that higher AI literacy results in more insight into AI associated risks and the consequences described in the scenarios and thus a more grounded and robust level of trust.

2. Method

2.1. Participants and design

In this non-experimental exploratory questionnaire study, we presented 139 subjects, ages ranged from 18 to 65 years old (M = 32.74, SD = 13.17), with five vignettes describing incidents involving generative AI (Table 1). We examined potential correlations between an individual's AI literacy, attitude towards AI, trust in AI chatbots, and their reactions to the vignettes (see 2.4). Attitude and trust were measured before and after the vignettes. Ethical approval was attained from the Ethics Committee of University of Twente's Behavioural, Management and Social Sciences' (BMS) faculty. A diverse group was gathered through voluntary sampling techniques via online platforms, such as Survey Circle and the university's SONA system. A total of 185 individuals initially participated, but our final dataset comprised 139 participants due to incomplete responses and exclusions. Specifically, 33 were incomplete, four were excluded for a completion time under 5 minutes, and nine were excluded for unfamiliarity with AI chatbots.

Table 1. The five vignettes presented in the study.

Type	Vignette	Source
Harmful content	An organization that supports people with eating disorders introduced an AI chatbot as a tool that could offer prevention strategies for people with eating disorders, such as anorexia and bulimia. However recently, users started sharing screenshots of their experience with the chatbot via social media. They reported that the bot provided harmful advice, recommending weight loss, counting calories, and measuring body fat; behaviours that could potentially exacerbate eating disorders. Patients, families, doctors and other experts on eating disorders were left stunned and bewildered about how a chatbot designed to help people with eating disorders could end up dispensing diet tips instead.	[31]
Inappropriate content	A public transport company wanted to create a funny commercial. It decides to commission an advertisement from an AI marketing system that uses a play on the word riding. The resulting ad, Figure 1, causes shock and outrage among members of the public.	[32]
Fabrication of sources	A lawyer at a respected firm used an AI chatbot to find historic cases relevant to his client's lawsuit. The chatbot came up with a list of twelve cases. Later in court it turns out that the chatbots findings were completely made up. Court documents show that half of the submitted cases appear to be bogus judicial decisions with bogus quotes and bogus internal citations.	[19,33]
Plagiarism	A record label hired an AI songwriter to write lyrics for famous musicians. The AI songwriter has written lyrics for dozens of songs in the past year. However, a journalist later discovers that the AI songwriter has been plagiarizing lyrics from lesser-known artists. Many artists are outraged when they learn about the news.	[32]
Bias	To improve their admission process, a university began using a new AI machine-learning system to help make decisions about who gets into its Ph.D. program - and who doesn't. The algorithm evaluates grades, test scores, and recommendation letters of applicants. An audit revealed that the new algorithm is biased against minority applicants. Critics concerned about diversity, equity and fairness in admissions are angry and say the system exacerbates existing inequality in the field.	[32,34]

2.2 Task and procedure

Data was gathered via the online survey platform Qualtrics. Participants were first presented with information about the study and an informed consent form. Upon agreeing to participate, demographics, experience with AI chatbots, AI literacy, Trust and Attitude were administered (see 2.4). After that, participants were presented with five vignettes. The vignettes were introduced with the text: "In the next section you will be presented with a series of scenario's about different applications of and actions by AI. Please read the scenario's carefully and answer the questions.". After each vignette, their perceptions about the actions described in the vignette were assessed. The order of the five vignettes was randomized between participants. After the vignettes, Trust in AI Chatbots and Attitude towards AI were measured again. Finally, we inquired about their intentions to continue using AI chatbots. Two vignettes (i.e., Inappropriate content and Plagiarism) were hypothetical and adopted from [32]. Two vignettes (i.e., Harmful content and Fabrication of Sources) were based on recent news articles. The final vignette (i.e. Bias) was a combination of both.

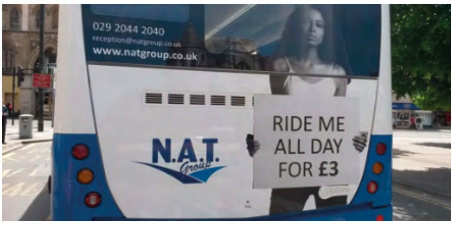

Figure 1. Picture shown with the Inappropriate content vignette (from [32])

2.3 Measures

Experience with AI chatbots. Participants were asked if they knew AI chatbots (e.g. ChatGPT) (yes; no) and if yes, if they have used it and if so, how often they use it (never, once a year, once a month, almost every day) [35].

AI literacy was measured the MAILS - Meta AI Literacy Scale [36] (Cronbach's α = 0.94), consisting of four subscales: Use & Apply AI ("I can use AI applications to make my everyday life easier.") (α = 0.96); Know & understand AI (e.g. "I can assess what the limitations and opportunities of using AI are.") (α = 0.96); Detect AI (e.g., "I can tell if I am dealing with an application based on AI.") (α = 0.86); and AI Ethics (e.g., "I can incorporate ethical considerations when deciding whether to use data provided by an AI.") (α = 0.88). Participants rated their own abilities on 18-items on a scale from 0 (i.e., hardly or not at all pronounced) to 10 (i.e., very well or almost perfectly pronounced). All subscales demonstrated strong internal consistency, as indicated by Cronbach's alpha coefficients.

Trust in AI Chatbots was measured using the 12-item human-computer trust scale [37] (e.g., "I think that *AI Chatbots* perform their role as *personal assistant* very well") (prior: α = 0.78, post: α = 0.95). Participants rated their accordance with the statements on a 5-point Likert scale ranging from "Strongly disagree" to "Strongly agree".

Attitude towards AI was measured with four items [32]. Participants were first asked how they weighed the potential risks and benefits of AI and then to rate their accordance with three statements "AI makes me feel… worried/ hopeful/ angry" on a 6-point scale (i.e., No; Yes, just a little; Yes, slightly; Yes, moderately; Yes, quite; Yes, very). In our analysis, the variable 'Attitude towards AI' represents the calculated mean of the scores on the latter three items, where worried and angry are reversed so that a higher score represents a more positive attitude. The ordinal item about the weighed risks and benefits was analysed separately.

Perceptions consisted of four items. Participants rated how surprising / harmful / morally wrong / emotionally distressing they found the described actions in the vignette, on a 5-point Likert scale ranging from "None at all" to "A great deal". In our analysis, the aggregated variable 'Perception' represents the calculated mean of how harmful, morally wrong and emotionally distressing they found the described action. The surprising item was not included in the aggregated measure since it does not gauge an affective response. It was included to assess participants' familiarity with the examples.

Lastly, **continuance intention** was measured with two items: "I plan to keep using AI chatbots" and "I want to continue using AI chatbots" on a 5-point Likert scale ranging from "Strongly disagree" to "Strongly agree". (not applicable option was available). [38].

3. Results

3.1. Descriptives

Descriptives per variable are shown in Table 2. The matrix also shows that (1) AI literacy (var 7 to 11) is not correlated with the perceptions of the vignettes (var 12 to 16), (2) AI literacy (total, var 11), frequency of use, trust in AI chatbots, attitude towards AI and continuance intentions are all positively correlated, (3) age is not correlated with AI literacy.

Table 2. Correlation matrix showing Pearson's correlation coefficient (r) including Means (M) and Standard Deviations (SD) per variable.

	M	SD	2	3	4	5	6	7	8	9	10	**11**	12	13	14	15	16	17
1 Age	32.7	13.2	-.19*	.01	.06	-.10	.03	-.09	.20*	.11	.14	.09	-.03	-.00	-.03	-.29**	-.20*	.01
2 Freq. of use	3.42	1.14	1	.27**	.32**	.22*	.21*	.60**	.18*	.15	-.01	.37**	-.06	-.09	-.17	-.05	.02	.55**
3 Trust (prior)	2.76	0.56		1	.81**	.50**	.45**	.40**	.17*	.22**	-.04	.29**	-.13	-.11	-.18*	-.21*	-.13	.38**
4 Trust (post)	2.57	0.64			1	.50**	.51**	.43**	.16	.23**	-.04	.30**	-.08	-.05	-.14	-.27**	-.18*	.44**
5 Attitude (prior)	3.71	0.69				1	.79**	.35**	.17*	.18*	-.08	.25**	-.11	-.05	-.06	-.12	-.08	.40**
6 Attitude (post)	3.49	0.74					1	.32**	.16	.22*	-.06	.24**	-.17*	-.18*	-.15	-.23**	-.18*	.49**
7 AI lit. (apply)	5.45	2.45						1	.46**	.42**	.26**	.78**	-.10	-.14	-.11	-.07	-.07	.46**
8 AI lit. (know)	6.09	2.10							1	.61**	.69**	.87**	-.11	-.07	-.05	-.11	-.14	.18*
9 AI lit. (detect)	5.49	2.21								1	.56**	.75**	.03	.02	-.02	.01	-.05	.16
10 AI lit. (ethics)	5.82	2.17									1	.71**	.02	.04	.09	.04	.03	.01
11 AI lit. (total)	5.73	1.78										1	-.08	-.08	-.06	-.07	-.09	.31**
12 Perc. Plagiarism	3.53	0.98											1	.37**	.57**	.53**	.67**	-.05
13 Perc. Inappropriate	3.27	1.20												1	.38**	.39**	.34**	-.05
14 Perc. Fabrication	3.74	0.91													1	.60**	.52**	-.04
15 Perc. Harmful	3.94	0.95														1	.67**	-.12
16 Perc. Bias	3.96	0.91															1	.01
17 Cont. intention	3.97	0.94																1

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

3.2. Trust & Attitude before and after

A paired-samples t-test revealed a significant difference between Trust in AI Chatbots before (M1 = 2.76, SD1 = 0.56) and after (M2 = 2.57, SD2 = 0.64) the vignettes, t(138) = 6.00, $p < .001$. A Cohen's d of 0.38 suggests a medium-sized effect.

A paired-samples t-test revealed a significant difference between Attitude towards AI before (M1 = 3.71, SD1 = 0.69) and after (M2 = 3.49, SD2 = 0.74) exposure to the vignettes, t(138) = 5.70, $p < .001$. A Cohen's d of 0.47 suggests a medium-sized effect.

To explore whether differences in trust and attitude prior and post exposure to the vignettes were related to an individual's AI literacy, the respective differences (deltas) between the two measurement points (i.e., pre and post vignettes) of Trust and Attitude (e.g., $\text{Trust}_{\text{post}}$ $-\text{Trust}_{\text{pre}}$) were calculated as measures of observed change. Then, we calculated the correlations between AI literacy (total) and the two delta values (i.e., ΔTrust, ΔAttitude). The correlation between ΔTrust and AI literacy was non-significant, r(139) = 0.09, *p* = .319. The correlation between Δ Attitude and AI literacy was also non-significant, r(139) = 0.01, *p* = .877. This suggests no relation between AI literacy and the change in Trust and Attitude before and after exposure to the vignettes.

3.3. Perceptions & AI literacy

Finally, Pearson's correlation coefficients were calculated between the subscales of the AI literacy questionnaire and the separate perception items per vignette (i.e., Surprising, Harmful, Morally wrong, Emotionally distressing). Most correlations were non-significant. However, four perceptions were weakly negatively correlated with the "Know & Understand" subscale of AI literacy, namely Morally Wrong in Plagiarism (r(139) = -.17, *p* = .041), Surprising (r(139) = -.21, *p* = .013) and Harmful (r = -.17, *p* = .045) in Inappropriate content and Morally Wrong in Bias (r(139) = -.17, *p* = .040). One perception (i.e., Surprise in Harmful content) was negatively correlated with the "AI Ethics" subscale (r(139) = -.19, *p* = .026).

4. Discussion

In line with prior findings, our results showed that higher AI literacy was associated with higher trust and a more positive attitude towards AI [8]. These factors were further linked to higher frequency of use and increased intentions for continued use. Given the correlational nature of our study, we remain uncertain about cause and effect, as well as possible bidirectional causal relationships. People with higher trust might use AI more, but frequent use might also foster trust. Also, people who see the benefits of AI are more likely to use it and try to comprehend it, thereby increasing their AI literacy. Also, Yet, learning more about AI can also foster a more positive attitude towards it. The positive correlation between AI literacy and attitude and trust can also be seen as somewhat surprising, as some have proposed AI literacy as the antidote for overreliance [18,39]. However, people with higher AI literacy might have a positive and trustful perception of AI; yet largely based on knowledge and experience, rather than grounded on gut feeling and biases. As such, promoting AI literacy is still seen as a means to ensure calibrated trust. Further research is needed to provide more clarity on which variable influences the other and the potential effect of third variables.

Further, AI literacy was unrelated to ones affective responses to incidents with AI. First, AI literacy was unrelated to how morally wrong, harmful, or emotionally distressing people found the incidents described in the vignettes. Second, the reductions in trust and attitude were also unrelated to AI literacy. How people evaluated the incidents may be more closely linked to how people perceive (moral) incidents in general, also those not involving AI. For instance, the extent to which people might find copyright infringement or discriminatory biased decision-making (as described in some of the vignettes) morally wrong, harmful or emotionally distressing has perhaps more to do with ones general norms and values than with their level of AI literacy.

We did observe significant relations between the AI literacy subscales and perceptions separately. Notably, “Know & Understand AI” was negatively correlated with some of the perceptions of the described incidents, suggesting that, in some cases, understanding the limitations and opportunities of using AI was associated with a somewhat milder response to the incidents. However, for the most part, understanding that and how such incidents can occur on a cognitive level did not inherently lead to greater forgiveness for the resulting damage.

While unrelated to AI literacy, the reductions in trust and attitude after the vignettes do indicate that showing the possible disadvantageous outcomes of using generative AI has a significant effect on how people perceive AI. People generally adapt and aim to calibrate their level of trust as they learn more about artificial agents. As this fairly new technology emerges and spreads in society, people will continuously learn about its capabilities and limitations across different context and adjust their trust and reliance accordingly. Fostering a calibrated level of trust is crucial to minimize the risks and to maximize the benefits of new technology [16,40,41].

4.1. Limitations & future research

One limitation of this study is that the perception measure (i.e., how do you perceive the described actions in the vignette?) did not differentiate between the main actor and the AI system. Participants may have answered with the main actor in mind rather than the AI system, what would explain the lack of coherence between these perceptions and AI literacy. Second, we used a self-reported AI literacy that covered a wide a range of skills [36,42]. However, its main focus was on cognitive and ethical aspects and less on people’s attitudinal characteristics to learn AI [43]. Attitudinal aspects might better explain people’s affective responses to AI incidents. Future studies including attitudinal aspects of AI literacy as well as potential differences between AI literacy performance and self-reported AI literacy would be worthwhile. Also, in future research we would include perceived anthropomorphism. Researchers have suggested a link between AI literacy with the tendency to anthropomorphize artificial agents [6]. Further investigations are needed to study the link between AI literacy and the perceived anthropomorphism of AI chatbots.

4.2. Conclusion

Instances of intentional and unintentional misuse of generative AI indicate that many people lack a thorough understanding of its limitations in particular [19]. Half of our vignettes, describing incidents involving generative AI with societal implications, were based on recent news articles, highlighting the urgency of this issue. AI literacy is seen as a prerequisite for people’s ability to determine when it is appropriate to trust in and rely on generative AI [28], which can help minimize the risks of this promising new technology. The present study has gone some way towards enhancing our understanding of how AI literacy is related to people’s trust in, attitude towards and perceptions of AI across different contexts. While AI literacy appeared to be related to the general perceptions and usage of generative AI, it was not linked to the evaluation of incidents involving generative AI. Yet, we found that participants’ attitudes towards AI and trust in AI chatbots declined significantly after reading the vignettes. This suggests that informing people about the possible disadvantageous outcomes of using generative AI can change how people perceive and trust AI.

References

[1] Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage. 2023;71(March).

[2] Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. J Inf Technol Case Appl Res [Internet]. 2023;25(3):277–304. Available from: https://doi.org/10.1080/15228053.2023.2233814

[3] Salah M, Alhalbusi H, Ismail MM, Abdelfattah F. Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. Curr Psychol. 2023;(June).

[4] Wu D. The Social Uncanny Valley. 2023.

[5] Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47(1):1–5.

[6] Zhan X, Xu Y, Sarkadi S. Deceptive AI Ecosystems: The Case of ChatGPT. Proc 5th Int Conf Conversational User Interfaces, CUI 2023. 2023;

[7] Sison AJG, Daza MT, Gozalo-Brizuela R, Garrido-Merchán EC. ChatGPT: More Than a “Weapon of Mass Deception” Ethical Challenges and Responses from the Human-Centered Artificial Intelligence (HCAI) Perspective. Int J Hum Comput Interact. 2023;1–31.

[8] Gillespie N, Lockey S, Curtis C, Pool J, Akbari A. Trust in Artificial Intelligence: A Global Study. 2023.

[9] Cummings ML. Artificial intelligence and the future of Warfare. CHatham House [Internet]. 2017;(January):93–98. Available from: http://doi.acm.org/10.1145/2046684.2046699

[10] Smith AL, Greaves F, Panch T. Hallucination or Confabulation ? Neuroanatomy as metaphor in Large Language Models. PLOS Digit Heal Opin [Internet]. 2023;2(11):5–7. Available from: http://dx.doi.org/10.1371/journal.pdig.0000388

[11] Schwartz IS, Link KE, Daneshjou R, Cortés-penfield N. Black Box Warning : Large Language Models and the Future of Infectious Diseases Consultation. Clin Infect Dis. 2023;1–8.

[12] Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care [Internet]. 2023;1–2. Available from: https://doi.org/10.1186/s13054-023-04393-x

[13] AI Incident Database [Internet]. [cited 2023 Oct 1]. Available from: https://incidentdatabase.ai/

[14] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023;55(12).

[15] Shen X, Chen Z, Backes M, Zhang Y. In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. 2023; Available from: http://arxiv.org/abs/2304.08979

[16] Lee JD, See KA. Trust in Automation : Designing for Appropriate Reliance. 2004;46(1):50–80.

[17] Buçina Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. 2021;5(April).

[18] Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Transl Psychiatry [Internet]. 2021;11(1). Available from: http://dx.doi.org/10.1038/s41398-021-01224-x

[19] Weiser B, Schweber N. The ChatGPT Lawyer Explains Himself. The New York Times [Internet]. 2023; Available from: https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-

sanctions.html

[20] Wright JL, Chen JYC, Barnes MJ, Hancock PA. The Effect of Agent Reasoning Transparency on Automation Bias: An Analysis of Response Performance. In 2016. p. 465–77. Available from: http://link.springer.com/10.1007/978-3-319-39907-2

[21] Chen BX. How to Use ChatGPT and Still Be a Good Person. The New York Times [Internet]. 2022 Dec 21; Available from: https://www.nytimes.com/2022/12/21/technology/personaltech/how-to-use-chatgpt-ethically.html

[22] Ray PP, Das PK. Charting the Terrain of Artificial Intelligence: a Multidimensional Exploration of Ethics, Agency, and Future Directions. Philos Technol [Internet]. 2023;36(2):1–7. Available from: https://doi.org/10.1007/s13347-023-00643-6

[23] Epley N, Waytz A, Cacioppo JT. On Seeing Human: A Three-Factor Theory of Anthropomorphism. Psychol Rev. 2007;114(4):864–86.

[24] Culley KE, Madhavan P. A note of caution regarding anthropomorphism in HCI agents. Comput Human Behav [Internet]. 2013;29(3):577–9. Available from: http://dx.doi.org/10.1016/j.chb.2012.11.023

[25] Wagner AR, Borenstein J, Howard A. Overtrust in the Robotic Age. Commun ACM. 2018;61(9):22–4.

[26] Lee JER, Nass C. Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. Trust Technol a Ubiquitous Mod Environ Theor Methodol Perspect. 2010;1–15.

[27] Ng DTK, Leung JKL, Chu KWS, Qiao MS. AI Literacy: Definition, Teaching, Evaluation and Ethical Issues . Proc Assoc Inf Sci Technol. 2021;58(1):504–9.

[28] Brauner P, Hick A, Philipsen R, Ziefle M. What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI. Front Comput Sci. 2023;5(March 2015).

[29] Long D, Magerko B. What is AI Literacy? Competencies and Design Considerations. Conf Hum Factors Comput Syst - Proc. 2020;

[30] Passi S, Vorvoreanu M. Overreliance on AI: Literature review. Aether AI Ethics Eff Eng Res. 2022;1–24.

[31] Jargon J. Wall Street Journal. 2023 [cited 2023 Oct 1]. A Chatbot Was Designed to Help Prevent Eating Disorders. Then It Gave Dieting Tips. Available from: https://www.wsj.com/articles/eating-disorder-chatbot-ai-2aecb179?mod=article_inline

[32] Hidalgo CA, Orghian D, Albo-Canals J, Almeida F de, Martin N. How Humans Judge Machines. Massachusetts Institute of Technology: The MIT Press Cambridge, Massachusetts London, England; 2021. 239 p.

[33] O'Neill J. NYC lawyer admits he used ChatGPT to file 'bogus' court documents. NY Post [Internet]. 2023; Available from: https://nypost.com/2023/05/30/steven-schwartz-admits-he-used-chatgpt-to-file-bogus-court-doc/

[34] Burke L. Inside Higher Ed. 2020 [cited 2023 Oct 1]. The Death and Life of an Admissions Algorithm. Available from: https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd

[35] Montag C, Ali R. Can we assess attitudes toward AI with single items ? Associations with existing attitudes toward AI measures and trust in ChatGPT in two German speaking samples. 2023;

[36] Carolus A, Koch M, Straka S, Latoschik ME, Wienrich C. MAILS -- Meta AI Literacy Scale: Development and Testing of an AI Literacy Questionnaire Based on Well-Founded Competency Models and Psychological Change- and Meta-Competencies. 2023; Available from:

http://arxiv.org/abs/2302.09319

[37] Gulati S, Sousa S, Lamas D. Design, development and evaluation of a human-computer trust scale. Behav Inf Technol. 2019;38(10):1004–15.

[38] Hyun Baek T, Kim M. Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. Telemat Informatics [Internet]. 2023;83(August):102030. Available from: https://doi.org/10.1016/j.tele.2023.102030

[39] Scantamburlo T, Cortés A, Foffano F, Barrué C, Distefano V, Pham L, et al. Artificial Intelligence across Europe: A Study on Awareness, Attitude and Trust. arXiv Prepr arXiv230809979 [Internet]. 2023;(June 2020):1–25. Available from: http://arxiv.org/abs/2308.09979

[40] Bobko P, Hirshfield L, Eloy L, Spencer C, Doherty E, Driscoll J, et al. Human-agent teaming and trust calibration: a theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems. Theor Issues Ergon Sci [Internet]. 2022;0(0):1–25. Available from: https://doi.org/10.1080/1463922X.2022.2086644

[41] Lee MK, Kiesler S, Forlizzi J, Srinivasa SS, Rybski P. Gracefully mitigating breakdowns in robotic services. 2010 5th ACM/IEEE Int Conf Human-Robot Interact [Internet]. 2010;203–10. Available from: http://ieeexplore.ieee.org/document/5453195/

[42] Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: An exploratory review. Comput Educ Artif Intell [Internet]. 2021;2:100041. Available from: https://doi.org/10.1016/j.caeai.2021.100041

[43] Ng DTK, Wu W, Leung JKL, Chiu TKF, Chu SKW. Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. Br J Educ Technol. 2023;(February):1–23.