AI-Related Misdirection Awareness In AIVR

Nadisha-Marie Aliman
Department of Information and Computing Sciences
Utrecht University
Utrecht, Netherlands
n.aliman@uu.nl

Leon Kester
Intelligent Autonomous Systems
TNO Netherlands
The Hague, Netherlands
leon.kester@tno.nl

***Abstract*—Recent AI progress led to a boost in beneficial applications from multiple research areas including VR. Simultaneously, in this newly unfolding deepfake era, ethically and security-relevant disagreements arose in the scientific community regarding the epistemic capabilities of present-day AI. However, given what is at stake, one can postulate that for a responsible approach, prior to engaging in a rigorous epistemic assessment of AI, humans may profit from a self-questioning strategy, an examination and calibration of the experience of their own *epistemic agency* – especially to counteract both intentional misdirection by unethical actors and unintentional epistemic self-sabotage. In this paper, we expound on a new avenue of utilizing AIVR tools to advance an AI-related *misdirection* awareness of humans in the deepfake era. Firstly, we harness scientific knowledge from the *psychology and neuroscience of magic* where the study of misdirection techniques is center stage. Secondly, we connect the latter to *creativity research* linking human creative potential to inspiration from the seemingly impossible. Overall, AIVR could become an empowering experiential testbed for human epistemic agency enabling a better rational evaluation of AI capabilities. However, a misuse of the same type of tools could yield AIVR safety risks if not counteracted preemptively.**

***Index Terms*—Misdirection, AIVR Ethics, AIVR Safety**

I. Introduction

A vital requirement for a democratic society which became highly salient in the deepfake era, is epistemic security [1], the protection of a society's knowledge creation and reasoning processes. Epistemic agency [2]–[5] relates to the experience of being *able* to actively contribute to such processes of participatory knowledge creation in the first place. Clearly, epistemic security would be at danger in a society where the majority of people do not own a sense of epistemic agency when exposed to malicious epistemic distortions including influence campaigns [5]. While both AI and VR technologies could plausibly open up tremendously valuable opportunities for humanity [6]–[8], one must proactively consider the multifarious socio-psycho-techno-physical harm that the use and misuse of same tools could engender [9]–[12] – including harm to human processes of knowledge creation and reasoning [13]. For instance, following Coeckelbergh [2], the human use of AI can threaten democracy *"since it risks diminishing the epistemic agency of citizens and thereby undermine the relevant kind of political agency in democracy"* [2]. Currently, humanity faces multifaceted disinformation and misinformation threats [14], [15] exacerbated by deepfake phenomena [16], [17] in various image, video, audio and text formats. Such epistemic threats may also analogously affect VR frameworks [18], [19]. More generally, VR godmother Nonny De La Peña [20] remarked in 2017 that VR could be misused for propaganda ends [21]. In line with this, Brown et al. [9] consider *"the unique threat of misinformation spread in VR to be real and relevant"* [9]. Moreover, in connection with AIVR-related epistemic security, researchers thematized the risk that malicious applications of conversational AI in immersive environments [13] could diminish human epistemic agency [5] if not thwarted adequately. On the whole, according to VR godfather Jaron Lanier, *"the more sophisticated technology becomes, the more damage we can do with it, and the more we have a "responsibility to sanity""* [22]. Indeed, from an ethical standpoint, one could argue that humans have the responsibility to stay vigilant and maintain or at least (re)gain epistemic agency in the face of critical issues impeding the control of technology. With AIVR having already been described to open up training and education opportunities for epistemic security [13], this paper investigates a new complementary AIVR avenue for supporting specifically an AI-related misdirection awareness to strengthen human epistemic agency.

Remarkably, psychological and neuroscientific research identified *"common factors in how people experience magic during a performance and are subject to misinformation"* [23] with magic being an object of scientific study [24]–[26] since more than a century [23] and culminating in the nascent field of the science of magic [27], [28]. More specifically, following Kuhn [28], the science of magic is to a large extent a scientific study of robust *misdirection* methods. Against this background, given the negative impact of disinformation and misinformation on epistemic agency in the deepfake era (be it through human-generated, AI-generated or hybrid artefacts), it could be advantageous to harness scientific knowledge from the psychology and neuroscience of magic to develop efficient countermeasures and foster human critical thinking. Interestingly, Alan Turing has been reported to have himself been inspired by a magic trick such as the Mechanical Turk [28], [29] to reason about "thinking machines". In the deepfake era, humans seem to be embedded in an ongoing kaleidoscopic imitation game with human-built AI. The human act of building more and more advanced imitative AI such as large language models can (appear to) be paired with a misdirection aspect [30] – which could be of benevolent or neutral origin and analogous to the magician proceeding as "honest liar".

However, next to the potential of powerful misdirection-based malicious exploits by unethical human actors, one must also avoid unconscious acts of epistemic self-sabotage where the use of human-built AI leads to humans fooling themselves. Relatedly, nowadays, even the information ecosystem of the scientific community did not stay immune to the totality of concerning AI-related epistemic threats[1] [37], [38]. In particular, authorship issues in the context of deepfake text emerged [37], [39] while more targeted scientific and empirical adversarial AI attacks are technically feasible [34] including the problem of AI-generated peer-review [40], AI-generated papers [41] and AI-generated empirical material [42].

On top of that, a few months ago, a divided, epistemically-relevant AI safety debate became highly salient in the computer science domain and beyond. Thereby, the underlying diverging epistemic assumptions on AI systems reflect fundamentally different approaches that can range from focusing on the short-term socio-technological risks of a sophisticated but manageable AI tool to instead foregrounding the long- or medium-term risk of a hypothetical human-built but uncontrollable AGI or artifical superintelligence that could ultimately qualitatively surpass humans in all tasks of interest. The goal of this paper is not to review all current epistemically-relevant AI assessments and there obviously exist many more nuances of conjectures on the same topic including diverging reflections on AI consciousness [43], [44]; instead, it is the presence of AI-related epistemic diagreements in general that we aim to highlight as it reflects the importance of both epistemic agency and the necessity of vigilance against misdirection and epistemic self-sabotage. Anecdotally, very recently, Gary Marcus remarked that the language model ChatGPT went from *"being mistaken for an AGI"* [45] to becoming part of a joke. Following Jaron Lanier, *"the danger isn't that a new alien entity will speak through our technology and take over and destroy us [...] the danger is that we'll use our technology to become mutually unintelligible [...] in a way that we aren't acting with enough understanding and self-interest to survive, and we die through insanity, essentially"* [22]. In this vein, this paper collates new ideas on how one could use AIVR to foster human critical thinking and creativity to avoid an unnecessary epistemic stagnation and relinquishment to a defeatist mode of epistemic self-sabotage. In Section II, we discuss how one could use knowledge from the psychology and neuroscience of magic to conceptually design an epistemic agency training in VR devised as a serious game that could foster AI-related misdirection awareness. Then, Section III explains how concepts from creativity research could be harnessed to stimulate human creativity in VR by questioning and interacting with the seemingly impossible – the latter also playing a role in magic misdirection. Finally, Section IV wraps up.

[1]Still, a modern philosophy of science based on and extending Karl Popper's [31] critical rationalism [32], [33] suggests that an explanation-anchored science can stay *resilient* [34] despite such attacks being possible since the process of creatively and disruptively discovering better yet unknown new scientific explanations of the world *cannot* be forged/imitated [35] (neither by people nor by present-day AI [36]). Instead of imitation, it involves unforeseeable, seemingly impossible transformation (see Section III).

II. VR For AI-Related Misdirection Awareness

Misdirection is at the core of magic tricks [46]. Beyond that, misdirection techniques can intrinsically underlie other forms of epistemic distortion. For instance, as can be extracted from the neuroscience and psychology of magic, it can inherently underlie misinformation mechanisms via attention illusions, memory illusions and illusions of choice *"in which we believe we are making decisions freely, but specific results are inevitable and out of our control"* [23]. Not suprisingly, magic misdirection methods have not only attracted research interest in domains such as VR [47], [48], game design [49], HCI [50], AI [51] and creativity research [52] but also in the context of information operations [53] and cyber operations [54]. Based on the psychological processing types that are key to human perception and cognition, Kuhn et al. [55] developed a taxonomy of misdirection comprising three different but interrelated processes: 1) perception, 2) memory and 3) reasoning. Following Kuhn, *"a magician can prevent a spectator from discovering the method by simply manipulating any one of these processes"* [28]. For an AI-related misdirection awareness training for humans in VR which promotes human epistemic agency, it would be valuable to apply a cybersecurity-oriented mindset [18] since defenders can profit from an adversarial perspective to craft ever better defenses. Concerning *perceptual misdirection*, VR seems to already be permeated by it as the very design of many immersive experiences and environments imply its application. Indeed, according to Bakk, one can interpret immersion in VR as a magical experiment [47]. In addition, the science of magic has already even explicitly been used to craft more vivid or convincing experiences in VR [28], [47], [48], [56] with Derren Brown's VR ghost train [57] being a notorious example. With regard to *memory misdirection*, researchers concluded that a variety of memory manipulations are technically feasible in VR [58]. Finally, concerning *reasoning misdirection* in VR, it may for instance play a role in scenarios of disinformation in VR [9], [18]. VR-based serious games have already been described to offer valuable interactive training avenues in various areas for creating safety and security awareness [59]–[61] and promoting experiential learning [62], [63]. In the following, we build on the misdirection taxonomy of Kuhn and colleagues [55] to exemplify how a serious game format in VR could support an AI-related misdirection awareness.

Firstly, perceptual misdirection includes attentional misdirection and non-attentional misdirection [55]. A few examples of perceptual misdirection in magic tricks are for instance attentional misdirection by affectively influencing the eye gaze of spectators via *"humor or other emotive content to misdirect the audience"* [23] and non-attentional misdirection by masking, a method of physically obstructing the view of a spectator such that the happening of a certain event is concealed [28]. Secondly, memory misdirection [55] can be mapped to forgetting and misremembering. Examples from magic could include forgetting because the magician utilizes *"techniques to prevent you from remembering certain details*

of an event" [28] or misremembering via a mechanism of *"recasting events that took place onstage in a manner that will bias spectators' memories of the performance"* [23]. Thirdly, reasoning misdirection relates to ruse, feigning and misguided assumptions about misdirection itself [28]. In a VR-based serious game to promote human AI-related epistemic agency, one could use generative AI to simulate these seven mentioned misdirection methods (attentional and non-attentional misdirection, forgetting, misremembering, ruse, feigning and misguided assumptions about misdirection). For this purpose, one could sample from a pool of available generative AI processes [18] (such as e.g. synthetic text generation, speech synthesis, adversarial perturbation but also VR deepfakes [18], [64]) that could be misused by malicious actors to cause epistemic distortions in AIVR. In particular, the integration of virtual conversational AI agents in a VR environment [5], [13] (abbreviated with VCAI in the following) may be useful to simulate present-day AI-related variants of misdirection processes. If present-day AI can be designed to in principle imitate anything that is imitable, then both from a safety-related and an ethical perspective, it becomes crucial to investigate and if possible reinforce the conjectured *non*-imitable aspects of people. In this respect, one could design a VR-based serious game with two epistemic levels (to be described in the next paragraph): an *imitation* versus a *transformation* level. The core idea is that while AI-based misdirection techniques including VCAIs can lead to indistinguishability in the imitation level, human users could explore and experience how by contrast, the transformation level could indeed be made robust to AI-aided but also human-crafted and hybrid misdirection strategies. Whether a form of distinguishability could be achieved in the transformation level would be contingent on the willingness and resoluteness of the participating people – which may foster their epistemic agency.

To design such a two-level epistemic serious game, designers need to select two qualitatively different tasks. The first task would be characterized by ease of forgery and high susceptibility to success by AI-related misdirection while the second task could be made highly robust to those characteristics. By way of illustration, for the imitation-centered level, one could take inspiration from the often misrepresented [65] initial idea of an imitation game by Alan Turing [66]. In his thought experiment on the imitation game, a human had to discern in a blind setting which of two participants was a woman [66]. Simply put, the participants initially mentioned were a man (A) and a woman (B) but the twist was to remplace A by a machine – leading to the two participants of the imitation game ultimately being a machine and a woman. More specifically, Turing asked: *"What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?"* [66]. In brief, the first imitation-centered part of the VR serious game could e.g. become a modern version of *"Find Turing's woman"* that is acceptable for the zeitgeist. Thereby, one could make use of a VCAI supported by any further suitable combination of generative AI processes to simulate misdirection. Concerning the second task for the transformation-centered level of the serious game which must be in principle amenable to robustness to misdirection, one could turn to science as use case. Science is historically linked to the non-imitative capability of people to perform disruptive acts of transformative creativity [67]–[69]. In this process, humans are challenging old theories of plausible utility and subsequently not only creating but also competently selecting unexpected new candidates of implausible utility [70] for better novel explanations of the world or even the discovery of new orders in nature [71]. Such creative acts can transform both people and the world they co-create (think for instance of the transformative effects that relativity and quantum theory had on people's view of the world). In sum, it is conceivable that a scientist is in principle able to meet the requirements for the transformation-centered level of the serious game.

Humans have been described as possessing extended minds [72] or extended conscious minds [73] and not to be limited to the resources of their biological body including the brain (although already the complexity of the latter may be often underestimated [74]). Already language can be understood as a technology [75]. In this sense, a human can be viewed as an inherently extended entity spanning a dynamic directed graph called a cyborgnet [36], [76] (a generic, substrate-independent and hybrid hierarchical unit [13], [36], [76] where explanatory narratives combine at least one entity able to consciously create and understand explanatory information (such as e.g. humans) and at least one entity that does not (such as e.g. present-day language AI, language itself, stone tools, fishes and so forth)). For instance, while remaining in charge of evaluating the generated outputs [38], [77] and retaining epistemic agency, scientists can apply language AI to broaden their divergent creativity [77] and generate adversarial outputs [76] to challenge their assumptions or for purposes of cognitive stimulation [78]. Scientists can e.g. use knowledge graphs [79] to deepen convergent creativity by unlocking tacit knowledge [80], i.e. the unknown known [81] (also called "dark matter of the mind" [82]) that is already *implied* by currently available knowledge. This may support the scientific endeavor which Popper [31] described as a process of bold novel conjectures followed by refutations via better new explanations [32]. Like language itself, language AI can be an augmentative tool within the scientist's cyborgnet. Thus, in the transformation-centered level, one could e.g. frame the game as a quest denoted: *"Find Popper's scientist cyborgnet"*. For a simple illustration, in this part of the serious game, the goal of a human evaluator could e.g. be to determine whether the set of three transacting entities A, B and C can be mapped to the cyborgnet of a scientist. In one VR room, a human scientist would be A while in the other VR room, a VCAI would take the part of A. In both rooms, B and C could be VCAIs able to fulfil the described assistive functions for scientists. Questions could be: *Which of the two VR rooms contains a cyborgnet of a scientist? Which generic strategies could reliably increase the distinguishability of the scientist cyborgnet?*

III. VR For AI-Related (Im)possibility Exploration

Instead of focusing on present-day AI's ability to achieve *in*distinguishability in an immersive imitation-centered setting (as is performed in the first part of the serious game), the questions in the transformation-centered part of the proposed open-ended serious game would concern the *distinguishability* of cyborgnets (in this case taking the example of human scientists). Importantly, the introduced idea of a VR-based serious game for AI-related misdirection awareness must feature a real-time interaction mode. In the transformation example from Section II, in one VR room, it would be required that a human scientist participates in real time and *not* via a generative AI tool acting as simulacrum [83] trained on past contents generated by that individual [81], [84]. Concurrently, precisely such simulacrum AI tools [81], [84] could be harnessed as strategy for the VCAI taking the part of A in the other VR room. One of the challenging but possible and efficient strategies a scientist could utilize to increase the own distinguishability in the presence of a human evaluator in this epistemic serious game, would be to create new previously unknown epistemic artefacts of implausible utility in real time. A strong candidate for such a strategy would be by generating *"creative leaps into the impossible"* [67]. Through the latter, the human evaluator could undergo a transformative experience easing distinguishability. As recently expounded by Corazza, humans *"can realize endeavors that are deemed to be impossible based on the shared extant knowledge at a certain time epoch"* [67] and *"can imagine impossible worlds, that are clearly out of the adjacent possible, and use these dreams to narrate fantastic stories, or to create games in virtual reality"* [67]. Corazza adds that a notorious example illustrating the usefulness of the impossibility element is Leonardo da Vinci *"who was able to imagine and describe ideas that where absolutely impossible in his epoch, and that were turned into reality up to four centuries later"* [67]. Indeed, discontinuous innovation and creativity is a highly relevant feature of science which has been depicted to be unfortunately often undervalued in the present scientific ecosystem [68], [69]. Overall, the VR-based epistemic serious game from the last Section II AIVR could be actively used to support humans in AI-related impossibility exploration and creativity in the service of strengthening human epistemic agency in the deepfake era.

While VR may be particularly suited for designing impossibility-focused experiences [85], [86], in the following, we briefly thematize the link to the scientific study of magic – with magic being an artform associated with the experience of the seemingly impossible [87] which can also be augmented by the use of AI tools [88]. In the literature, a connection between impossible experiences and creativity has been postulated [52]. Thereby, experiencing the seemingly impossible via magic tricks [28], [52] (next to experiencing it in VR [89] or via dream imagery [52]) may represent a possible creativity-enhancing avenue. In their study, Wiseman and Watt concluded that *"a considerable amount of research across a diverse range of contexts suggests that experiencing the impossible promotes creative and expansive thinking"* [52]. Beyond that, a recent study corroborated a link between the experience of magic performances and subsequently enhanced divergent creativity [24]. Finally, another study found *"surprise of an unexpected, impossible moment to be driving the enjoyment in magic"* [87]. Moreover, following Morgan and colleagues, *"magic is far more than a technique (but the use of knowledge and insights to create an experience with an audience)"* [68]. Interestingly, since the very design of the VCAIs for the two VR-based serious game levels from Section II are conceptually grounded in magic misdirection methods, it appears plausible that the scientist could derive creative inspiration and even enjoyment from the functioning of those very tools crafted to either imitate a person or distract away from a scientist cyborgnet (e.g. via distractive narratives and storytelling [23] for attentional misdirection). Given the creativity-stimulating avenues of VR itself (which as expounded in a recent neuroscientific study is able to positively affect creative processes in the brain [90]), one can conjecture that both evaluator and scientist cyborgnet – if willing to – can self-empower to co-create distinguishability from VCAIs via transformative one-shot moments of experiencing the seemingly impossible.

IV. Conclusion and Future Work

In this paper, we motivate a new complementary research avenue of relevance for AIVR ethics and safety that aims at supporting the urgently required epistemic agency of humans in the deepfake era. We expounded on how one could conceptually design a bipartite VR-based serious game with virtual conversational AI agents (VCAIs) to forward an *AI-related misdirection awareness* enabling a better rational assessment of AI capabilities. For this purpose, we built on scientific knowledge from the *psychology and neuroscience of magic* where the study of misdirection techniques is center stage. Moreover, we established a link to *creativity research* according to which humans gain useful inspiration from the seemingly impossible. Given the epistemic security threat of unintelligibility through AI-related indistinguishability, the two-level epistemic serious game that we described leads from the exploration of misdirection-susceptible *imitation* to more robust creative *transformation* strategies. In light of multifarious AI(VR) safety risks including those linked to future VCAIs [5], it seems vital *not* to *solely* focus on the short-term heuristic of AI detection that may be insufficient in the long-term [91]–[94] – also in view of unceasing adversarial cat and mouse games [95]–[101] and the possible unnecessary stigmatization of human statistical outliers [34], [92]. Instead, once epistemic agency is regained, it is possible that technology-augmented critical thinking can be sharpened to the point of achieving cyborgnetic distinguishability via the non-imitative disruptive creativity moments that could benefit humanity. Future work could extend such open-ended AIVR safety approaches by scientific knowledge from the study of other transformative art forms [102], [103]. In short, people could harness AIVR to counteract the epistemic threat of what Jaron Lanier referred to as death "through insanity" [22].

References

[1] E. Seger, S. Avin, G. Pearson, M. Briers, S. Ó. Heigeartaigh, H. Bacon, H. Ajder, C. Alderson, F. Anderson, J. Baddeley *et al.*, "Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world," *The Alan Turing Institute*, 2020.

[2] M. Coeckelbergh, "Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence," *AI and Ethics*, pp. 1–10, 2022.

[3] H. Gunn, M. P. Lynch, and J. Lackey, "The internet and epistemic agency," *Applied epistemology*, pp. 389–409, 2021.

[4] E. Miller, E. Manz, R. Russ, D. Stroupe, and L. Berland, "Addressing the epistemic elephant in the room: Epistemic agency and the next generation science standards," *Journal of Research in Science Teaching*, vol. 55, no. 7, pp. 1053–1075, 2018.

[5] L. Rosenberg, "The Metaverse and Conversational AI as a Threat Vector for Targeted Influence," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2023, pp. 0504–0510.

[6] J. Cowls, A. Tsamados, M. Taddeo, and L. Floridi, "A definition, benchmark and database of AI for social good initiatives," *Nature Machine Intelligence*, vol. 3, no. 2, pp. 111–115, 2021.

[7] D. M. Markowitz and J. N. Bailenson, "Virtual reality and the psychology of climate change," *Current Opinion in Psychology*, vol. 42, pp. 60–65, 2021.

[8] N. Tomašev, J. Cornebise, F. Hutter, S. Mohamed, A. Picciariello, B. Connelly, D. C. Belgrave, D. Ezer, F. C. v. d. Haert, F. Mugisha *et al.*, "AI for social good: unlocking the opportunity for positive impact," *Nature Communications*, vol. 11, no. 1, p. 2468, 2020.

[9] J. Brown, J. Bailenson, and J. Hancock, "Misinformation in Virtual Reality," *Journal of Online Trust and Safety*, vol. 1, no. 5, 2023.

[10] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[11] K. Hartmann and K. Giles, "The next generation of cyber-enabled information warfare," in *2020 12th International Conference on Cyber Conflict (CyCon)*, vol. 1300. IEEE, 2020, pp. 233–250.

[12] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, "Risks and benefits of large language models for the environment," *Environmental Science & Technology*, vol. 57, no. 9, pp. 3464–3466, 2023.

[13] N.-M. Aliman and L. Kester, "VR, Deepfakes and Epistemic Security," in *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2022, pp. 93–98.

[14] V. Bufacchi, "Truth, lies and tweets: A consensus theory of post-truth," *Philosophy & Social Criticism*, vol. 47, no. 3, pp. 347–361, 2021.

[15] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[16] E. Horvitz, "On the Horizon: Interactive and Compositional Deepfakes," *arXiv preprint arXiv:2209.01714*, 2022.

[17] N. Schick, *Deep fakes and the infocalypse: What you urgently need to know*. Hachette UK, 2020.

[18] N.-M. Aliman and L. Kester, "Malicious design in AIVR, falsehood and cybersecurity-oriented immersive defenses," in *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2020, pp. 130–137.

[19] H. Wu, P. Hui, and P. Zhou, "Deepfake in the Metaverse: An Outlook Survey," *arXiv preprint arXiv:2306.07011*, 2023.

[20] L. Knoepp, "Forget Oculus Rift, Meet The Godmother Of VR," https://www.forbes.com/sites/lillyknoepp/2017/04/13/forget-oculus-rift-meet-the-godmother-of-vr/, 2017, Forbes; accessed 04-August-2020.

[21] QUARTZ, "Virtual reality, fake news and the future of fact," https://www.youtube.com/watch?v=i5LW03vw_x8, 2017, YouTube video; accessed 04-August-2020.

[22] S. Hattenstone, "Tech guru Jaron Lanier: 'The danger isn't that AI destroys us. It's that it drives us insane'," https://www.theguardian.com/technology/2023/mar/23/tech-guru-jaron-lanier-the-danger-isnt-that-ai-destroys-us-its-that-it-drives-us-insane, 2023, The Guardian; accessed 26-August-2023.

[23] R. G. Alexander, S. L. Macknik, and S. Martinez-Conde, "What the Neuroscience and Psychology of Magic Reveal about Misinformation," *Publications*, vol. 10, no. 4, p. 33, 2022.

[24] T. Li, L. E. McCalla, H. Zheng, and Y. Lin, "Exploring the influence of magic performance on design creativity," *Thinking Skills and Creativity*, vol. 47, p. 101223, 2023.

[25] S. Macknik, S. Martinez-Conde, and S. Blakeslee, *Sleights of mind: What the neuroscience of magic reveals about our everyday deceptions*. Henry Holt and Company, 2010.

[26] R. Q. Quiroga, "Magic and cognitive neuroscience," *Current Biology*, vol. 26, no. 10, pp. R390–R394, 2016.

[27] G. Kuhn, A. A. Amlani, and R. A. Rensink, "Towards a science of magic," *Trends in cognitive sciences*, vol. 12, no. 9, pp. 349–354, 2008.

[28] G. Kuhn, *Experiencing the impossible: The science of magic*. MIT Press, 2019.

[29] T. Standage, *The Turk: The life and times of the famous eighteenth-century chess-playing machine*. Walker & Company, 2002.

[30] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks, "AI Deception: A Survey of Examples, Risks, and Potential Solutions," *arXiv preprint arXiv:2308.14752*, 2023.

[31] K. Popper, *Conjectures and refutations: The growth of scientific knowledge*. Routledge, 1963.

[32] D. Frederick, *Against the Philosophical Tide: Essays in Popperian Critical Rationalism*. Critias Publishing, 2020.

[33] ——, "Falsificationism and the Pragmatic Problem of Induction," *Organon F*, vol. 27, no. 4, pp. 494–503, 2020.

[34] N. M. Aliman and L. Kester, "Epistemic defenses against scientific and empirical adversarial AI attacks," in *CEUR Workshop Proceedings*, vol. 2916. CEUR WS, 2021.

[35] D. Deutsch, *The beginning of infinity: Explanations that transform the world*. penguin uK, 2011.

[36] N.-M. Aliman, *Cyborgnetics – The Type I vs. Type II Split*. Kester, Nadisha-Marie, 2021.

[37] A. Birhane, A. Kasirzadeh, D. Leslie, and S. Wachter, "Science in the age of large language models," *Nature Reviews Physics*, pp. 1–4, 2023.

[38] D. Leslie, "Does the sun rise for ChatGPT? Scientific discovery in the age of generative AI," *AI and Ethics*, pp. 1–6, 2023.

[39] H. Y. Jabotinsky and R. Sarel, "Co-Authoring with an AI? Ethical dilemmas and artificial intelligence," *Ethical Dilemmas and Artificial Intelligence (December 15, 2022)*, 2022.

[40] M. Hosseini and S. P. Horbach, "Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other Large Language Models in scholarly peer review," *Research Integrity and Peer Review*, vol. 8, no. 1, p. 4, 2023.

[41] A. Jungherr, "Using ChatGPT and Other Large Language Model (LLM) Applications for Academic Paper Assignments," *Otto-Friedrich-Universität*, 2023.

[42] L. Wang, L. Zhou, W. Yang, and R. Yu, "Deepfakes: a new threat to image fabrication in scientific publications?" *Patterns*, vol. 3, no. 5, 2022.

[43] P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji *et al.*, "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," *arXiv preprint arXiv:2308.08708*, 2023.

[44] J. Kleiner and T. Ludwig, "If consciousness is dynamically relevant, artificial intelligence isn't conscious," *arXiv preprint arXiv:2304.05077*, 2023.

[45] G. Marcus, "The rise and fall of ChatGPT?" https://garymarcus.substack.com/p/the-rise-and-fall-of-chatgpt, 2023, Homepage; accessed 26-August-2023.

[46] G. Kuhn, P. Kingori, and K. P. Grietens, "Misdirection–Magic, Psychology and its application," *Science & Technology Studies*, vol. 35, no. 2, pp. 13–29, 2022.

[47] Á. K. Bakk, "Magic and immersion in VR," in *Interactive Storytelling: 13th International Conference on Interactive Digital Storytelling, ICIDS 2020, Bournemouth, UK, November 3–6, 2020, Proceedings 13*. Springer, 2020, pp. 327–331.

[48] S. Marwecki, A. D. Wilson, E. Ofek, M. Gonzalez Franco, and C. Holz, "Mise-Unseen: Using eye tracking to hide virtual reality scene changes in plain sight," in *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 2019, pp. 777–789.

[49] S. Kumari, S. Deterding, and G. Kuhn, "Why game designers should study magic," in *Proceedings of the 13th International Conference on the Foundations of Digital Games*, 2018, pp. 1–8.

[50] B. Tognazzini, "Principles, techniques, and ethics of stage magic and their application to human interface design," in *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, 1993, pp. 355–362.

[51] W. Smith, F. Dignum, and L. Sonenberg, "The construction of impossibility: a logic-based analysis of conjuring tricks," *Frontiers in psychology*, vol. 7, p. 748, 2016.

[52] R. Wiseman and C. Watt, "Experiencing the impossible and creativity: a targeted literature review," *PeerJ*, vol. 10, p. e13755, 2022.

[53] K. Scott, "'Nothing up my sleeve': Information warfare and the magical mindset," in *Cyber Influence and Cognitive Threats*. Elsevier, 2020, pp. 53–76.

[54] S. Henderson, R. Hoffman, L. Bunch, and J. Bradshaw, "Applying the principles of magic and the concepts of macrocognition to counter-deception in cyber operations," in *the12th International Meeting on Naturalistic Decision Making, McLean, VA*, 2015.

[55] G. Kuhn, H. A. Caffaratti, R. Teszka, and R. A. Rensink, "A psychologically-based taxonomy of misdirection," *Frontiers in psychology*, vol. 5, p. 1392, 2014.

[56] C. Maraffi, "Stage Magic as a Performative Design Principle for VR Storytelling," *Maraffi, C.(2021). Stage Magic as a Performative Design Principle for VR Storytelling. Cinergie–Il Cinema E Le Altre Arti*, vol. 10, no. 19, pp. 93–104, 2021.

[57] R. Manthrope, "Derren Brown's VR ghost train is back– and this time it's actually scary." http://www.wired.co.uk/article/derren-brown-vr-ghost-train-thorpe-park, 2017, WIRED; accessed 26-August-2023.

[58] E. Bonnail, W.-J. Tseng, M. McGill, E. Lecolinet, S. Huron, and J. Gugenheimer, "Memory Manipulations in Extended Reality," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–20.

[59] Z. Feng, V. A. González, C. Mutch, R. Amor, A. Rahouti, A. Baghouz, N. Li, and G. Cabrera-Guerrero, "Towards a customizable immersive virtual reality serious game for earthquake emergency training," *Advanced Engineering Informatics*, vol. 46, p. 101134, 2020.

[60] R. Lovreglio, D.-C. Ngassa, A. Rahouti, D. Paes, Z. Feng, and A. Shipman, "Prototyping and testing a virtual reality counterterrorism serious game for active shooting," *International Journal of Disaster Risk Reduction*, vol. 82, p. 103283, 2022.

[61] S. V. Veneruso, L. S. Ferro, A. Marrella, M. Mecella, and T. Catarci, "CyberVR: an interactive learning experience in virtual reality for cybersecurity related issues," in *Proceedings of the International Conference on Advanced Visual Interfaces*, 2020, pp. 1–8.

[62] E. A. Alrehaili and H. Al Osman, "A virtual reality role-playing serious game for experiential learning," *Interactive Learning Environments*, vol. 30, no. 5, pp. 922–935, 2022.

[63] S. S. Oyelere, N. Bouali, R. Kaliisa, G. Obaido, A. A. Yunusa, and E. R. Jimoh, "Exploring the trends of educational virtual reality games: a systematic review of empirical studies," *Smart Learning Environments*, vol. 7, pp. 1–22, 2020.

[64] J. T. Hancock and J. N. Bailenson, "The social impact of deepfakes," pp. 149–152, 2021.

[65] J. E. Larsson, "The Turing test misunderstood," *ACM SIGART Bulletin*, vol. 4, no. 4, p. 10, 1993.

[66] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[67] G. E. Corazza, "Beyond the adjacent possible: On the irreducibility of human creativity to biology and physics," *Possibility Studies & Society*, vol. 1, no. 1-2, pp. 37–45, 2023.

[68] R. M. Morgan, R. L. Kneebone, N. D. Pyenson, S. B. Sholts, W. Houstoun, B. Butler, and K. Chesters, "Regaining creativity in science: insights from conversation," *Royal Society Open Science*, vol. 10, no. 5, p. 230134, 2023.

[69] I. Yanai and M. J. Lercher, "Make science disruptive again," *Nature Biotechnology*, vol. 41, no. 4, pp. 450–451, 2023.

[70] J. Tsao, C. Ting, and C. Johnson, "Creative outcome as implausible utility," *Review of General Psychology*, vol. 23, no. 3, pp. 279–292, 2019.

[71] D. Bohm, "Quantum theory as an indication of a new order in physics. Part A. The development of new orders as shown through the history of physics," *Foundations of Physics*, vol. 1, pp. 359–381, 1971.

[72] A. Clark and D. Chalmers, "The extended mind," *analysis*, vol. 58, no. 1, pp. 7–19, 1998.

[73] P. Telakivi, "A Roadmap from the Extended Mind to the Extended Conscious Mind," in *Extending the Extended Mind: From Cognition to Consciousness*. Springer, 2023, pp. 1–31.

[74] A. Roli, J. Jaeger, and S. A. Kauffman, "How organisms come to know the world: fundamental limits on artificial general intelligence," *Frontiers in Ecology and Evolution*, vol. 9, p. 1035, 2022.

[75] D. Everett, *How language began: The story of humanity's greatest invention*. Profile Books, 2017.

[76] N.-M. Aliman and L. Kester, *Immoral programming: What can be done if malicious actors use language AI to launch 'deepfake science attacks'?* Wageningen Academic Publishers, 01 2022, pp. 179–200.

[77] J. M. Buriak, D. Akinwande, N. Artzi, C. J. Brinker, C. Burrows, W. C. Chan, C. Chen, X. Chen, M. Chhowalla, L. Chi *et al.*, "Best Practices for Using AI When Writing Scientific Manuscripts: Caution, Care, and Consideration: Creative Science Depends on It," pp. 4091–4093, 2023.

[78] T. McIntosh, T. Liu, T. Susnjak, H. Alavizadeh, A. Ng, R. Nowrozy, and P. Watters, "Harnessing GPT-4 for Generation of Cybersecurity GRC Policies: A Focus on Ransomware Attack Mitigation," *Computers & Security*, p. 103424, 2023.

[79] Y. Hou, J. Yeung, H. Xu, C. Su, F. Wang, and R. Zhang, "From Answers to Insights: Unveiling the Strengths and Limitations of ChatGPT and Biomedical Knowledge Graphs," *medRxiv*, pp. 2023–06, 2023.

[80] J. Senker, "The contribution of tacit knowledge to innovation," *Cognition, Communication and Interaction: Transdisciplinary Perspectives on Interactive Technology*, pp. 376–392, 2008.

[81] D. Van Buren, "Guided scenarios with simulated expert personae: a remarkable strategy to perform cognitive work," *arXiv preprint arXiv:2306.03104*, 2023.

[82] D. L. Everett, *Dark matter of the mind: the culturally articulated unconscious*. University of Chicago Press, 2019.

[83] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," *arXiv preprint arXiv:2304.03442*, 2023.

[84] E. Schwitzgebel, D. Schwitzgebel, and A. Strasser, "Creating a large language model of a philosopher," *arXiv preprint arXiv:2302.01339*, 2023.

[85] S. M. Ritter, R. I. Damian, D. K. Simonton, R. B. van Baaren, M. Strick, J. Derks, and A. Dijksterhuis, "Diversifying experiences enhance cognitive flexibility," *Journal of experimental social psychology*, vol. 48, no. 4, pp. 961–964, 2012.

[86] C. Velasco, F. Barbosa Escobar, O. Petit, and Q. J. Wang, "Impossible (food) experiences in extended reality," *Frontiers in Computer Science*, vol. 3, p. 716846, 2021.

[87] S. E. Bagienski and G. Kuhn, "A balanced view of impossible aesthetics: An empirical investigation of how impossibility relates to our enjoyment of magic tricks," *i-Perception*, vol. 14, no. 1, p. 20416695221142537, 2023.

[88] H. Williams and P. W. McOwan, "Magic in pieces: An analysis of magic trick construction using artificial intelligence as a design aid," *Applied artificial intelligence*, vol. 30, no. 1, pp. 16–28, 2016.

[89] T. Siimon, K. Tulver, K. K. Kaup, M. Vasser, and J. Aru, "Facilitating real-life creative insight through psychedelic virtual reality," 2023.

[90] Y.-Y. Wang, T.-H. Weng, I.-F. Tsai, J.-Y. Kao, and Y.-S. Chang, "Effects of virtual reality on creativity performance and perceived immersion: A study of brain waves," *British Journal of Educational Technology*, vol. 54, no. 2, pp. 581–602, 2023.

[91] W. Antoun, V. Mouilleron, B. Sagot, and D. Seddah, "Towards a Robust Detection of Language Model Generated Text: Is ChatGPT that Easy to Detect?" *arXiv preprint arXiv:2306.05871*, 2023.

[92] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, "GPT detectors are biased against non-native English writers," *arXiv preprint arXiv:2304.02819*, 2023.

[93] J. Otterbacher, "Why technical solutions for detecting AI-generated content in research and education are insufficient," *Patterns*, vol. 4, no. 7, 2023.

[94] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-generated text be reliably detected?" *arXiv preprint arXiv:2303.11156*, 2023.

[95] M. Campbell and M. Jovanović, "Detecting Artificial Intelligence: A New Cyberarms Race Begins," *Computer*, vol. 56, no. 8, pp. 100–105, 2023.

[96] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors

to adversarial examples," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3348–3357.

[97] Z. Jiang, J. Zhang, and N. Z. Gong, "Evading Watermark based Detection of AI-Generated Content," *arXiv preprint arXiv:2305.03807*, 2023.

[98] C. Liu, H. Chen, T. Zhu, J. Zhang, and W. Zhou, "Making DeepFakes more spurious: evading deep face forgery detection via trace removal attack," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[99] N. Lu, S. Liu, R. He, and K. Tang, "Large Language Models can be Guided to Evade AI-Generated Text Detection," *arXiv preprint arXiv:2305.10847*, 2023.

[100] G. Martínez, L. Watson, P. Reviriego, J. A. Hernández, M. Juarez, and R. Sarkar, "Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet," *arXiv preprint arXiv:2306.06130*, 2023.

[101] V. Veselovsky, M. H. Ribeiro, and R. West, "Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks," *arXiv preprint arXiv:2306.07899*, 2023.

[102] M. D. Murgia, "Teaching synaesthesia as a gateway to creativity," *Exchanges: The Interdisciplinary Research Journal*, vol. 2, no. 2, pp. 305–313, 2015.

[103] S. Preminger, "Transformative art: art as means for long-term neurocognitive change," *Frontiers in human neuroscience*, vol. 6, p. 96, 2012.