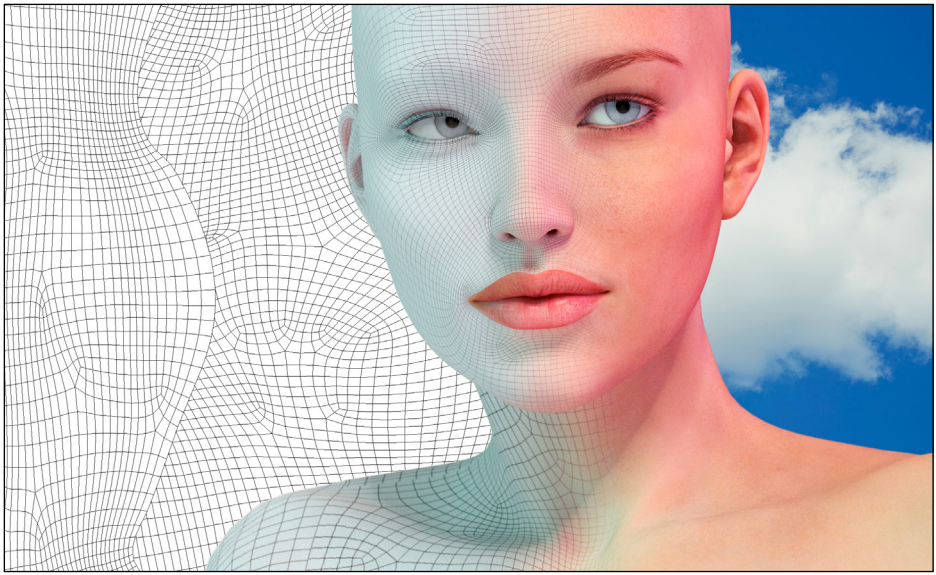
Climbing the uncanny valley

Solomon Gilbert



Key digested message

Much has been said of AI's risks in the hands of criminals, particularly new advances in generative AI. But should we be concerned? This article offers a perspective away from the prevailing opinion amongst experts and argues that we have less to be concerned about, than what is speculated. Fraud has been around since the dawn of the human condition. Many of our modern advances have contributed to its reach and impact, but it is doubtful that AI will be one of them.

Keywords: Artificial intelligence, Fraud, Online risks, Evolution, Online crime, Technological advancements.

'IS SOMETHING troubling you? What would you like to discuss? Tell me more…'

These are the words of ELIZA; the creation of Joseph Weizenbaum. Designed to mimic Rogerian psychotherapy, this 1967 program mirrored the complex prisms of the human experience simply by using a list of pre-made responses to 'reflect' an interactant's words in question form (Eliza, Computer Therapist, n.d.). Weizelbaum and ELIZA made a glancing blow to the rusting armour of our human instincts using a programming language called SLIP (Symmetric LIst Processor) (SLIP Programming Language, n.d.). As our first attempt at the Turing Test – the ability of machines to exhibit behaviour indistinguishable from humans – it performed with reasonable success. Accompanied by nothing but a pre-defined list of questions, ELIZA provoked an emotive, passionate

https://doi.org/10.53841/bpsadm.2024.16.1.57

response in participants' relationship to the machine (Jones & Bergen, 2023). Almost 60 years on, not much has changed.

Our yearning to relate to others is etched within the human psyche, and the aching need for a simple, understandable truth is the mechanism with which we leave ourselves vulnerable (Shang et al., 2022). As Weizenbaum later decried; convincing ourselves of ELIZA's anthropomorphic acuity is just as fundamentally wrong as ascribing it with humanity (Weizenbaum, 1976). It is, after all, a pre-fated program. If you read the source code, you already know what it's going to say. It is incapable of the empathetic expanse of the human experience. But its participants authentically believed their understanding of its emotions to be true (Weizenbaum, 1976). As recent as October 31st 2023, ELIZA out-performed GPT-3.5 in the Turing Test. More people were convinced that a human was responding from ELIZA's replies than they were by the 'latest-in-technology' GPT-3.5 (Jones & Bergen, 2023).

Crime targets human nature. Nothing is worth more to criminals than the time saved in recreating the human experience and its vulnerabilities. From the 419 scams of the past to its modernity, the business model of fraud is not to persuade everyone beyond doubt that they're taking to a real person, but to make it just believable enough that a desperate individual will be convinced (Konradt et al., 2016). While its believability has been pervasive for decades, the public consciousness has only recently been enamoured by AI's mimicry. With this admonishment comes the inevitable howl of 'are we safe?'

The short answer is yes. AI adds no more value to the business of fraud than any other tool. Criminality pursues adaptation. With the proliferation of AI, it has adapted different mechanisms to perpetuate the same nonsense, fraudulent claims (Case Study: The Cyber Security of Artificial Intelligence, 2023). What were the charismatic snake oil salesmen of the 18th century, are now fake AI-generated voices of celebrities shilling the very same useless products and medications. Nevertheless, the technological evolution of fraud is married to the same intrinsic evolution of our understanding of truth. Language adapts. The language of online communication adapts. The cliché misspellings of fake scam emails weren't always the butt of a joke. The language of what's true has always been a quintessential part of our world (Wells, 2018).

At the recent safety summit, many a faux word was waxed lyrical by the Prime Minister about the emphatic concern he has with AI's capabilities beyond rationale. The potential for AI to take over the jobs of illegally employed disparate workers in impoverished countries to defraud vulnerable individuals was absent from consideration. Instead, his valuable time was dedicated to ensuring that high skilled jobs weren't overtaken (Chair's summary of the AI Safety Summit, 2023). His apprehensions were misplaced. The greatest threat AI poses is not in the upper echelons of skilled work, but on the ground floor of script-driven scam call centres (Ferrera, 2023).

It's very easy to point to AI, certainly Large Language Models (LLMs), as a source of information. LLMs will assert what they 'believe' to be true, even if they're completely mistaken. These models source their understanding of truth from what information is available to them at the time. But the way AI language models are trained doesn't make them concerned with objectivity as much as they are being convincing purveyors of falsehood (OpenAI, 2023). To criminals, their accuracy isn't as important as persuasively rendering the human experience. AI is not a threat because it can teach us precisely how to make explosives or poisons – that information is already widely available on

the internet to those with ill intent. It's a threat because it can accurately impersonate the human condition (Herley, 2012). It's climbed the uncanny valley and convinced us of its legitimacy. As prosaic as it is to listen to someone recount a dream, it's equally bone-crunchingly boring to listen to someone narrate the (undoubtedly) transcendent experience they had with ChatGPT.

As the epiphanic experience of interacting with an LLM percolates through the real lives of people sharing ideas online, so too will their adjustment in their expectations of language. The lexicon of Chat Generative Pre-trained Transformer (ChatGPT, for those interested) may be powerful; AI generated faces and voices may be persuasive; but the nuance and poignancy of real communication will eventually sculpt itself around AI's injection of sham into our idea sharing methods. Our wisdom for truth will merely gain a more pinpoint understanding. This has been the case throughout history (Wells, 2018). We recognise the phony patter of a scam email because it's become a part of our collective consciousness. We adapt. It is within us to discern fact from fiction, and artificial intelligence is only true in its artifice, not in its intelligence.

The fact remains, however, that the mechanism of fraud is in convincing people of an enticing untruth, controlling their behaviour in such a way that benefits the fraudster. If AI can decisively pass the Turing Test, what is to stop it from targeting our most vulnerable? Back of the class: Weren't you paying attention? Criminals have had the ability to hook vulnerable people into an attractive lie since the invention of people and, by extension, vulnerability (Law and Order in Ancient Egypt. The Development of Criminal Justice from the Pharaonic New Kingdom Until the Roman Dominate, 2014).

Online fraud is a net with which our culprit fishermen catch their marks. Factors contributing to the rate and success of online crime cut this metaphor in two distinct halves: the size of the net; and the fineness of its mesh (Maguire, 2019). The net's size dictates how many individuals can be reached by the criminal. A large-scale email or social media campaign promoting a scam doesn't have to be convincing; it just must reach a large enough population that the probability is high enough that someone will believe it. The increasing ease at which information is spread; the ever-growing infrastructure of criminal groups; and our rapid interconnectedness all give criminals greater access to our population at large.

Our second, more complicated factor, is how fine they make their netting – how relevant, believable, or persuasive a fraud must be to a population (Maguire, 2019). If too many potential targets slip through the holes of a fraud's relevance, it becomes inefficient. If too many are caught, it risks overrunning the capacity of the perpetrator's resources.

Intricate societal undercurrents play a huge part in weaving this fabric. Confusing or contradictory messaging from authority makes it hard to discern truth, and instances of crime rise – as was the case during the shaky public health advice of the pandemic (Zhang et al. 2022). As the complexity of a system increases, so too does the rate of fraud – no more salient an example can be found than in the lead-up to the 2008 financial crash (Griffin, 2021), where financial tools were extremely complicated, and fraud was rife. Complex and confusing systems such as our world wide web allows crime to blossom. To conclude this metaphor; AI is the cool new colour of the net.

AI does harm. It harms minorities when used to decide employment prospects (Iriodo, 2018). It harms lithium miners in developing countries (CFR, 2022). It harms our environment with its increasing demand on processing power (Coleman, 2023). Its use

as a tool to commit crime will undoubtedly result in some success for our opposition, as did the printing press, the newspaper, and the email. But AI does not afford the criminal any more reach to their audience. It can't capitalise the raw emotions which drive people to fall victim. It can't cash in on the social confusion and complexity which makes fraud effective. Those skills are distinctly human, and I believe will remain so.

The beating heart of fraud is a claim. It may be a claim designed to elicit greed, fear, confusion, deception, or panic. The antidotes and advice against falling victim to fraud has never been focused on the vehicle by which the claim was made. Instead, the focus has always been in empowering people to discern truth. The 'take 5' initiative set by the NCSC advises people who may be targeted by scams to consciously evaluate claims made by criminals online (Take Five, n.d.). Armouring the population with questions such as 'is this too good to be true?' is far more effective than asking them to spot minute, ever-changing technical details in emails or social media chats. Douglas Rushkoff (2010) wrote 'We are looking at a society increasingly dependent on machines, yet decreasingly capable of making or even using them effectively'. Teaching the language of crime untethers society from their reliance on machines and throws water on the paper tiger of AI.

The author

Solomon Gilbert has always had a passion for curiosity and a burning desire to know. This curiosity, happily married with a yearning to press large red buttons he shouldn't, as Solomon expressed, got him into some hot water in the past. However, it has also given him the knowledge and capability to meaningfully help others. Solomon dedicated his working life to making the online world a safer place. Starting in cyber security at 17 years old, he spent the last 9 years providing companies with the guidance they need to protect themselves from fraud; and government bodies such as the NCA with the understanding they need to prevent cybercrime. Beyond his job of combating cyber enabled crime, the desire to serve his community extends into the time he spends sitting as a magistrate.

References

Case study: The cyber security of artificial intelligence. (n.d.). https://www.ncsc.gov.uk/collection/annual-review-2023/technology/case-study-cyber-security-ai

CFR Editors. (2022, June 28). Artificial intelligence's environmental costs and promise. Council on Foreign Relations. https://www.cfr.org/blog/artificial-intelligences-environmental-costs-and-promise

Chair's summary of the AI Safety Summit 2023, Bletchley Park. (2023, November 2). GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park

Coleman, J. (2023, November 30). AI's Climate Impact Goes beyond Its Emissions. Scientific American. https://www.scientificamerican.com/article/ais-climate-impact-goes-beyond-its-emissions/

Eliza, computer therapist. (n.d.). https://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm

Ferrara, E. (2023). GenAI against Humanity: Nefarious applications of generative artificial intelligence and large language models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.00737

Griffin, J.M. (2021). Ten years of evidence: Was fraud a force in the financial crisis? *Journal of Economic Literature, 59*(4), 1293–1321. https://doi.org/10.1257/jel.20201602

Herley, C. (2012). Why do Nigerian Scammers Say They are from Nigeria? ResearchGate. https://www.researchgate.net/publication/242070582_Why_do_Nigerian_Scammers_Say_They_are_from_Nigeria

Iriondo, R.I. (2018, October 11). Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women – Machine Learning – CMU – Carnegie Mellon University. Machine Learning | Carnegie Mellon University. https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html

Jones, C. & Bergen, B. (2023). Does GPT-4 pass the Turing Test? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.20216

Konradt, C., Schilling, A. & Werners, B. (2016). Phishing: An economic analysis of cybercrime perpetrators. *Computers & Security, 58*, 39–46. https://doi.org/10.1016/j.cose.2015.12.001

Law and Order in Ancient Egypt. The Development of Criminal Justice from the Pharaonic New Kingdom until the Roman Dominate. (2014, December 24). Student Repository. https://studenttheses.universiteitleiden.nl/handle/1887/30196

Maguire, E.R. (2019). Testing the fraud triangle: a systematic review. *Journal of Financial Crime, 27*(1), 172–187. https://doi.org/10.1108/jfc-12-2018-0136

OpenAI. (2023, March 23). GPT-4 System Card [Press release]. https://cdn.openai.com/papers/gpt-4-system-card.pdf

Rushkoff, D. (2010). *Program Or be Programmed: Ten Commands for a Digital Age.* OR Books.

Shang, Y., Wu, Z., Du, X., Jiang, Y., Ma, B. & Chi, M. (2022). The psychology of the internet fraud victimization of older adults: A systematic review. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.912242

SLIP (programming language). (n.d.). Academic Dictionaries and Encyclopedias. https://en-academic.com/dic.nsf/enwiki/2173195

Take Five. (n.d.). Take five – To stop fraud | To stop fraud. https://www.takefive-stopfraud.org.uk/

Weizenbaum, J. (1976). *Computer power and human reason: From Judgment to Calculation.* W.H. Freeman.

Wells, J.T., Riley, R., McNeal.A. & Holderness JR.K. (2018, August 1). How the evolution of language affects fraud risk. Journal of Accountancy. https://www.journalofaccountancy.com/issues/2018/aug/how-language-affects-fraud-risk.html

Zhang, Y., Wu, Q., Zhang, T. & Yang, L. (2022). Vulnerability and fraud: evidence from the Covid-19 pandemic. *Humanities and Social Sciences Communications, 9*(1). https://doi.org/10.1057/s41599-022-01445-5