Understanding the Interplay between Trust, Reliability, and Human Factors in the Age of Generative AI

Dr Simon Thorne

Department of Computer Science, Cardiff School of Technologies
Cardiff Metropolitan University, Wales, UK.

Email: sthorne@cardiffmet.ac.uk

***Abstract* - In the swiftly evolving landscape of Generative AI, particularly through Large Language Models (LLMs), there is a promising utility across diverse applications. While these tools promise heightened accuracy, efficiency, and productivity, the potential for misinformation and "hallucinations" underscores the need for cautious implementation. Despite proficiency in meeting user-specific demands, LLMs lack a comprehensive problem-solving intelligence and struggle with input uncertainty, leading to inaccuracies. This paper critically examines the nuanced challenges surrounding Generative AI, delving into trust issues, system reliability, and the impact of human factors on objective judgments. As we navigate the complex terrain of Generative AI, the presentation advocates for a discerning approach, emphasizing the necessity of verification and validation processes to ensure the accuracy and reliability of generated outputs. The exploration serves to illuminate the multifaceted dimensions of trust in technology, providing insights into how human factors shape our ability to make objective assessments of the reliability and accuracy of artefacts produced by Generative AI. This contribution to the academic discourse fosters a comprehensive understanding of the intricate dynamics inherent in the responsible utilisation of Generative AI technologies.**

Keywords - Generative AI, Trust in Technology, Human Factors, Software Engineering, Verification, Validation, Hallucinations

I. INTRODUCTION

This paper will explore LLMs and generative AI, what such systems are capable of in the hands of novice users. Major threats to the integrity of LLM produced artefacts are considered in the context of hallucinations and inaccuracies arising from uncertainty in the input prompt. The concept of trust in automation is explored through the lens of the user, how this trust can be misplaced, breached and the consequences of such failures. The context of trust in generative AI is finally considered, paying particular attention to the lack of professional experience and training in software engineering of the novice user, the limits of their knowledge and how that affects their ability to validate and verify code generated from LLMs.

II. GENERATIVE AI AND LLMS

"Generative AI", the process of using AI to generate text, computer code, video or images has advanced to a stage where relatively accurate responses can be gained from plain English user prompts. This empowers the user to create a range of different artefacts that can be leveraged in a variety of settings, for instance creating copy for social media or blog posting, creating Python computer code for data science projects or generating video presentations to advertise or supplement written materials.

Generative AI is based on Large Language Models which are deep learning neural networks that consume a vast corpus of human writing during the training phase of development. This corpus of text allows the neural network to determine the relationships and connections that exist between different words, sentences and paragraphs and as such predict the meaning of a passage of text or determine the underlying intent in a user provided "prompt" and generate a suitable output. Current examples of this technology provide relatively accurate responses to requests although this is dependant on the quality of the prompt provided. If the user is able to describe exactly what is required in the prompt with all relevant detail and nuance, then LLMs can provide a highly accurate response, whether that is generated text, computer code or still or moving pictures.

However, In certain circumstances, LLMs can provide incorrect information, misleading answers and "hallucinations" which distort accuracy and reliability of LLMs. Hence the user must take responsibility for validating and verifying the outputs of any generative AI model to ensure that the response is reliable and accurate. In short the user most not blindly trust the reliability of any LLM generated output.

A. Hallucinations and Inaccuracies

Hallucinations and subsequent distortions in truth or accuracy are the biggest barrier and threat to full exploitation of generative AI technology. Hallucinations are defined by IBM [1] as

"…a phenomenon wherein a large language model (LLM) perceives patterns or objects that are non-existant or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate."

Hallucinations are greatly influenced by the wording of the prompt given [1]. Using particular phrases or language can induce hallucination responses in LLMs, one such example considers GPT interpretation of Boolean Logic statements [2]. In this example, Boolean logic questions were posed to GPT, the questions were taken from a computer science class on logic. The

questions themselves test whether two different logic gates are equivalents in terms of processing input and output.

Every one of these questions starts with the phrase "Show that…" i.e. show that a 2 Input OR gate produces an output that is the same as a 2 input NAND gate. There are 10 questions in total, but with one intentionally impossible proposition designed to make the students hunt for the incorrect answer. When GPT was posed with these questions, it answered correctly, it was also able to provide all of the different logic gates and implications of NOT. However, when it encountered the intentionally impossible question, it answered positively, saying that the two gates were equivalent and provided reasoning and proof. This was initially confusing since GPT appeared to answer the other questions correctly, however the responses were being dictated by the phrase at the start of the prompt "show that". When this phrase was changed to ask a question, "Does a 2 input AND gate produce the same output as a 2 input OR gate", GPT answered negatively. LLMs are particularly good at meeting the intent of the user, in using the phrase "show that" the LLM was being instructed to show that these logic constructs were the same, regardless of whether that is true or not.

Other research shows that LLMs can struggle with different logical and mathematic reasoning such as the execution of BODMAS [3, 4] incorporating the implications of negation [5, 2] and activities such as inference and deduction especially where there is uncertainty in the prompt.

B. Reasoning with uncertainty

LLMs appear to struggle to reason with uncertainty or in situations where inference is needed to arrive at the correct conclusion. This inability to reason with uncertainty is evidenced through LLM performance in logic puzzles [6]. Research shows that in problems where all of the information needed to solve the puzzle is presented in the prompt, LLMs are capable of coming to the correct conclusion. However, if the puzzle contains the need to infer the correct answer through deduction, LLMs have a very limited ability to execute this correctly, often leading to hallucinations and inaccurate responses [6, 2, 7].

This specific limitation to LLMs could have serious implications for the reliability of artefacts generated by humans using this technology. The major problem is that whilst we might be able to get LLMs to produce exactly what we want with a very carefully crafted prompt, this level of effort requires insight into how LLMs work in order to avoid the pitfalls. A large number of LLM users will either be unaware of the limitations or will trust generative AI to a point where they believe that LLMs are a general problem solving intelligence, capable of deducing the correct answer in any circumstance. This use of LLMs comes from a few distinct processes, the first is the tendency for humans to trust computers and the output of computer systems over time, the second is the tendency for humans to use technology in ways not recommended or anticipated by the vendors and the third is the tendency for humans to invest "magical" properties in computer artefacts, especially where "Artificial Intelligence" is the driving force.

C. Trust in Technology and Automation

The concept of trust is critical in many domains that human beings work in. To trust an individual is to bestow an assumption of reliability and predictability of a co-operative agenda that is built up over time through interacting with an individual. Trust in technology and automation broadly observes the same principles to develop trust in automated systems over time.

Muir [9] defines trust as "An attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability" and that trust in automation is therefore the belief that automated systems will effectively assist users in achieving their objectives even in uncertain or vulnerable situations.

Users will place different level of trust in automated systems which can vary between a simple reliance on system functionality through to a more complex and nuanced approach to trust that is based on the reliability and performance of an automated system.

Users calibrate their trust in automated systems, this is the alignment between user trust levels and the actual performance of a system. This trust develops with the experience of the user, and it is essential to the safe and efficient operation of such systems. When a user's trust is correctly calibrated, informed decisions can be made about when it is safe to rely on the system or whether its outputs should be verified.

Trust is also dynamic, it adapts continuously based on the users experience with the system, the feedback received and varying task conditions. There are various dimensions of automated system performance that influence user trust. Reliability is the user perception of the systems reliability and performance in completing tasks. Transparency is the degree to which the systems operations, decision making processes and limitations are transparent to the user. The experience of the user over time, past interactions and feedback obtained is also important. Familiarity and training therefore also form a large part of calibrating trust, how familiar is the system to the user and what training has been delivered to establish the knowledge needed to operate the system in a variety of conditions. Finally, the complexity and the level of uncertainty in the task also affect a user's trust in the system, novel uncertain situations may lower the trust of the user in the system.

Muir [8, 9] based these observations on the operation of automated factory equipment but these ideas have been used as the basis for describing trust in computer systems and in human computer interaction more widely. Research into trust in automation and computer systems has taken these original ideas and developed them, a good overview of these models is given by [10].

If a computer system is perceived as trustworthy and the particular function of that system is important enough, it can become "reified" in the eyes of those who use it. Reification is where an abstract idea, like a computer application or data model is transformed by trust into a "real" object that becomes the unquestionable truth, users of those systems perceive the reified system as the embodiment of that idea and as such is bestowed with properties such as believability, correctness, appropriateness, concreteness, integrity, tangibility, objectivity and authority [11].

However, this trust can often be misplaced and the level of trust given to such a system can backfire on those depending on it.

D. Misplaced Trust

Unfortunately, there are countless examples of misplaced trust in computer systems which have varying consequences for organisations, governments and entire societies. Trust is misplaced in these examples through the assumption of reliability in the software or the analysis that informs a decision. Computer artefacts are used for the most critical of decision making, the following two examples show how flawed analysis or mistakes in process can be life and death issues.

In the wake of the financial crisis in 2008, two Harvard professors published a paper "Growth in a time of debt" [12] that performed an analysis on GDP to debt ratios for various countries to determine the optimal debt to GDP ratios. The paper asserted that once a country exceeded 90% debt to GDP ratio, the only outcome was negative growth. This conclusion was used by governments around the world to justify austerity politics, George Osbourne the chancellor in the UK at the time even made direct reference to the 90% limit in parliament. However, the analysis was flawed, several rows of data in the analysis spreadsheet had been accidentally omitted from the analysis, if these rows had been included, the only conclusion that could be reached is that there is no hard limit between debt and GDP ratios and hence the justification for austerity had no basis in fact.

There was never any correction of austerity politics or even acknowledgement that the evidence used to justify this course of action was false. Research shows that between 2012 and 2019, there were 335K excess deaths in the UK which have been attributed to austerity politics and a general increase in mortality across high income nations [13, 14].

In 2020 during the height of the global COVID pandemic, it was reported that the UK £42B test and trace system had 'lost' 16,000 positive test cases through a format issue in a spreadsheet. The test and trace system batch processed positive test cases from various authorities in the UK. This data was collected locally and then combined to produce overall statistics reported by the UK government and SAGE. It is unclear precisely what happened but either data was imported into an older format of Excel (.xls) most likely from a comma separated value (CSV) file or somewhere in the pipeline of data flow, there was an older version of Excel using the .xls format. This format could only represent 65,000 rows of data. When a file larger than this is encountered, the "bottom" part of file is simply chopped off, accounting for the 16,000 lost test cases. Those lost test cases meant that individuals with a positive diagnosis of COVID19 were unaware of their condition and as such were not taking relevant precautions to protect others. The number of missed contacts those infected individuals have had is thought to be 48,000 [15] who would have gone on to infect others and inevitably some of those individuals must have died as a result.

The above examples are picked to illustrate misplaced trust in computer systems. In the case of austerity, the analysis provided was trusted. The analysis was published in a leading journal following peer review, the contents of the paper were trusted by decision makers who used the evidence to justify serious economic decisions which have far reaching implications. In the case of COVID19, the software and data pipelines were trusted to be reliable and predictable since surveillance of the progression of the virus was a key part of the strategy to manage the COVID19 outbreak in the UK.

Given the importance of both tasks, the highest levels of scrutiny should have been applied before the analysis and software were both fully trusted to complete their tasks with precision and reliability. Clearly in either case, the operation of the system was not verified or validated sufficiently, the same can be said of the faulty analysis, basic checking would have highlighted this omission. Yet these systems were trusted to perform and they both catastrophically failed, inflicting serious consequences on a large number of people depending on them.

E. Trust and Generative AI

Generative AI presents some serious concerns through the accessibility of the tool, assuming plain English statements can be written for the prompt by the user, then the ability to generate a variety of outputs is possible. Some of these outputs do not require "validation" as such, for instance artwork or images. Interpretation of the image as appropriate or inline with the vision of the user is enough to confirm the validity of the artefact. Generating language or code is a different matter.

To validate language, the user must know the domain sufficiently to confirm that the text generated is accurate and appropriate. If generative AI is being used outside of the users experience, then validation of text in a foreign domain is not possible based on their existing knowledge. Hence unless the user develops a strategy for dealing with this, for instance triangulation of facts and assertions, then the trust in reliability of the information is blind.

Computer code presents an even more complex situation, whilst the code can be compiled to validate its function and operation, validation alone does not guarantee that the code really meets the needs of the user or the task at hand. For instance, consider that we create some software that can collect product reviews and classify them as positive or negative with generative AI.

We can generate and test the code to ensure that it meets the specification of the problem and executes as expected. Imagine then that the code was implemented but was classifying sarcastic product reviews as positive when quite the opposite is meant. Here only verification, checking the alignment between the computer system and the real world, could have prevented this bug in our system from subverting our goals.

There is also then the possibility that using generative AI to generate code is producing suboptimal code that may present security vulnerabilities [16] or through human verification be proved to be in accurate [17]. If the user does not possess the knowledge to understand that code must be verified then such subtleties are likely to be missed and bugs may be introduced which will become apparent when the right conditions to trigger them arise.

There is also significant risk the use of generative AI to produce computer code bypasses many well established software engineering principles and processes designed to improve and assure quality. For instance, a novice user with no training or experience in software engineering will not understand the need to plan the system in advance in order to minimise the chance of omission in the function and specification of the software. Testing processes may be missed through the same lack of knowledge in software engineering which may result in undetected errors, inefficiencies or other negative software outcomes. Hence using generative AI to produce code without sufficient knowledge and experience in software engineering presents a looming AI driven coding crisis that threatens to reduce quality and contribute to many more catastrophic software failures.

III. CONCLUSIONS

Generative AI is constantly adapting and represents a great opportunity for increased productivity and efficiency. It also offers us the opportunity to leverage knowledge and techniques outside of our expertise without the need to invest in training. Potentially, this is very valuable but we must be cautious with such power.

In circumstances where we are outside of our domain, knowledge or skillset, we risk providing inaccurate or suboptimal artefacts that we cannot validate or verify. In these situations we are unable to calibrate our trust in the system, we cannot objectively evaluate whether the output is trust worthy or not. In such circumstances, there is a danger that our own perceptions of generative AI could distort our perception through processes like “Magical Thinking” [18] which describes the tendency of humans to believe that “AI” systems possess capabilities or characteristics that are beyond their actual capabilities. For instance projecting human like qualities onto AI like a general problem solving ability, consciousness, intuition or the assumption that the output is always correct. This “Magical Thinking” is similar to reification but without the need for scrutiny of the artefact, magical thinkers bestow qualities on AI without the need to see the evidence sufficient to reach such conclusions.

In summary, generative AI will provide access to complex activities like software engineering to non software engineers. This will mean a net increase in software that has not been “designed” beforehand, that may contain inefficiencies, vulnerabilities or even hidden errors through a lack of real testing, validation and verification. In that vein, it has great similarities to the introduction of the spreadsheet which has afforded many non-computing professionals access to complex data processing. The world of spreadsheets is one of frequent errors and catastrophic failure [19], generative AI will be no different unless we can build a reliable mechanism to objectively evaluate the level of trust that should be placed in such systems.

REFERENCES

[1] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li and Y. Xia, "Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models," In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), p. 245–255, 2023.

[2] S. Thorne, "Experimenting with ChatGPT for spreadsheet formula production: Evidence of Risk in AI Generated Spreadsheets," in Proceedings of The European Spreadsheets Risks Interest Group annual conference, London, 2023.

[3] A. Borji, "A Categorical Archive of ChatGPT Failures," Preprint Arxiv, 2023.

[4] S. Frieder, L. Pinchetti, R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Peterson, A. Chevalier and J. Berner, "Mathematical Capabilities of ChatGPT," Preprint Arxiv, 2023.

[5] T. A. Chang and B. K. Bergen, "Language Model Behavior: A Comprehensive Survey.," Computational Linguistics, pp. 1-50, 2024.

[6] V. Plevris, G. Papazafeiropoulos and A. Jiménez Rios, "Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard," AI, vol. 4, pp. 949-969, 2023.

[7] Y. Wan, W. Wang, Y. Yang, Y. Yuan, J.-t. Huang, P. He, W. Jiao and M. R. Lyu, "A & B == B & A: Triggering Logical Reasoning Failures in Large Language Models," Arxiv, 2024.

[8] B. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," Ergonomics, vol. 37, p. 1905–1922, 1994.

[9] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," Ergonomics, vol. 39, p. 429–460, 1996.

[10] S. Kohn, E. de Visser, E. Wiese, T. Lee and T. Shaw, "Measurement of Trust in Automation: A Narrative Review and Reference Guide," Frontiers in Psychology, 2021.

[11] G. Croll, "The Reification of an Incorrect and Inappropriate," in Proceedings of the annual conference of The European Spreadsheets Risks Interest Group, London, 2017.

[12] C. Reinhardt and K. Rogoff, "Growth in a Time of Debt," AMERICAN ECONOMIC REVIEW, vol. 100, no. 2, pp. 573-578, 2010.

[13] D. Walsh, R. Dundas, G. McCartney, M. Gibson and R. Searman, "Bearing the burden of austerity: how do changing mortality rates in the UK compare between men

and women?," Journal of Epidemiol Community Health, vol. 76, pp. 1027-1033, 2022.

[14] G. McCartney, L. Fenton, J. Minton, C. Fischbacher, M. Taulbut, K. Little, C. Humphreys, A. Cumbers, F. Popham and R. McMaster, "Is austerity responsible for the recent change in mortality trends across high-income nations? A protocol for an observational study," Britsh Medicial Journal Open, no. 10, 2019.

[15] D. Gerard, "How Not to Kill People With Spreadsheets," Foreign Policy, October 2022.

[16] M. Taeb, H. Chi and Bernadin, "Assessing the Effectiveness and Security Implications of AI Code Generators," Journal of the Colloquium for Information Systems Security Education, 2024.

[17] J. Liu, C. Xia, Y. Wang and L. Zhang, "Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation," in Advances in Neural Information Processing Systems 36, 2023.

[18] M. Morris, "Magical thinking and the test of humanity: we have seen the danger of AI and it is us," AI and Society , 2023.

[19] R. Panko, "Spreadsheet errors: What we know and what we think we can do," in Proceedings of The European Spreadsheets Risks Interest Group annual conference, London, 2000.