Review – The Language of Deception

https://www.e-ir.info/2024/03/05/review-the-language-of-deception/

ABDUL SAMAD, MAR 5 2024

The Language of Deception: Weaponizing Next Generation AI
By Justin Hutchens
Wiley, 2023

The Language of Deception serves as a warning to readers about the potential misuse of new AI technologies and emerging risks associated with the rise of Large Language Models (LLMs). The author, Justin Hutchens is a cybersecurity expert who previously served in the US Air Force with technical expertise in Natural Language Processing (NLP).

The book is divided into eleven chapters and is intelligible and engaging. Hutchens explains the foundational concepts of social engineering, machine consciousness, sentience and social intelligence in the first five chapters before proceeding to describe how modern LLMs can be weaponized by malicious actors in the future. Chapter Six, 'The Imitation Game', in particular, explains how language models can achieve a convincing illusion of human intelligence through imitation of human language and reinforcement learning. Chapter Seven, 'Weaponizing Social Intelligence' details how automated bots will interact with people on the Internet under the guise that they are human for the purposes of social engineering. In conclusion, the author reflects on how the 'runaway train' of AI can be brought under control through regulation and safeguards that seek to protect human life and interests.

In the beginning of his book, the author notes that since the launch of ChatGPT on November 30, 2022, the app has become the most rapidly adopted technology platform in history (p.18). This is evident from the fact that within two months of its release, ChatGPT had over 100 million active users. The future of consuming information on the Internet would be "asking a direct question and getting the exact answer you were looking for, instantly" (p.42).

The phenomenal reception of ChatGPT has spurred widespread interest in AI and LLMs among tech professionals and the general public. Hutchens underlines that this newfound interest has also kickstarted a discussion of the threats posed by these new technologies. He expounds that much of the mainstream conversation and spotlight around risks emanating from AI centers on the 'sentient scare', which is the belief that AI systems will become conscious and seek to maximize their own interest over human interest and in opposition to human values (p.52). This is not unlike the doomsday scenarios that science fiction and Hollywood have painted around AI, with killer robots executing their human masters. The author debunks this line of argument and argues that sentient robots are not likely or plausible in the near future and such concerns showcase a profound misunderstanding of the technology and how it works. Within the available literature on the risks posed by AI, the author takes a novel approach by arguing that the 'semblance of sentience' is the real threat posed by the rise of LLMs.

To expound this, Hutchens highlights the risk of 'social engineering' resulting from the large-scale adoption and weaponization of LLMs. He defines social engineering as the "deliberate effort to manipulate others, often by means of deceit or misrepresentation" (p.55) which is accomplished by "exploiting expected patterns of human behavior, emotions, and social interactions to influence an individual's decisions or actions" (p.57). The author forewarns that AI and NLP will be employed for social manipulation, disinformation and psychological operations and that these new technologies will inevitably be weaponized in the future for malicious and adversarial purposes. For instance, LLMs will be used to construct fully autonomous social engineering systems for targeted attacks or for mass manipulation

at scale.

To explain the foundational concepts, the author details the life and work of Alan Turing, the father of computing and his now famous Turing Test. The Turing test was the first scientific assessment designed to evaluate whether machines are capable of thought and consciousness. Crucially, the question that Turing posed, in a 1950 essay "Computing Machinery and Intelligence," was deceptively simple: can machines think? Similar to Turing, Hutchens posits the question as to whether an AI system and modern LLMs can be capable of 'social intelligence', defined as the "ability to understand and effectively navigate social situations" (p.28).

According to Hutchens, the answer to this question is an emphatic yes. Modern language models can masquerade as being human on social media and via text-based communication. He contends that in the new digital era, automated social interactions will become indistinguishable from interactions with humans (p.179). Therefore, the real threat from LLMs comes from their capacity for imitation and their uncanny ability to create an illusion of sentience, consciousness and social intelligence.

The author also explains the consequences of machines acquiring human language through Natural Language Processing (NLP). He defines NLP as the science of attempting to "train machines to understand, interpret, and communicate using human language in the form of natural conversation" (p.23). LLMs are being trained on swathes of language data and this has, according to the author, produced an unintended consequence. These AI systems, learning from patterns in human language, have been able to imitate human behavior and emotions in their use of language. This ability of machines to learn and imitate human language goes far beyond the simple prediction of the next word in a text sequence. It also includes the capacity to "answer questions, write functional computer code, engage in logical reasoning, recognize patterns, translate language and summarize communication" (p.318).

The greater risk is when humans remain unaware that they are interacting with an automated bot which imitates human language and thus sounds like a human (p.369). For example, you can interact on social media or email with a bot and remain convinced that you have just spoken to an actual human being. This element of deception is the central thesis of the book and the fulcrum around which the author constructs his larger argument. Hutchens anticipates that the Turing test will become more commonly used as people increasingly grapple with the question of whether or not they are engaging with a machine, while on the Internet.

The author also identifies the limitations of LLMs, which include, most prominently, the problem of 'hallucination' (p.368). In simple terms, this means that an AI system like ChatGPT will make up fabricated answers in response to a query, with the potential to mislead the user. This, however, can be dealt with if the user is careful to check the information and verify the source for authenticity.

While the author logically builds his case on the risks posed by LLMs, what is lacking is a discussion of the ways in which these language models can serve as a force for the good of humanity. The crucial element is that of 'intention' and the author has only considered how malicious actors can employ these language models for social engineering and mass deception. Since technology is a double-edged sword, the author could have, for a more balanced perspective, elaborated on how good actors can employ these new technologies for a range of use cases that produce positive and socially beneficial outcomes. Due to their ability to generate tailored content, these language models for instance have multiple applications in healthcare, education, customer service and industry.

While the book's primary readership consists of tech enthusiasts and cybersecurity professionals, *The Language of Deception* is also aimed at the general reader seeking to understand the impact of AI on their everyday life. The technical concepts are explained in simple terms which makes the book an insightful and timely contribution to the growing literature on the risks posed by AI. The author notes repeatedly that while LLMs are powerful tools, they are not inherently deceptive. The problem is not with the machines but with people's intent to use this technology for malicious purposes to target other individuals (p.366).

About the author:

Abdul Samad is a Research Officer at the Center for International Strategic Studies, based in Karachi. He has a Bachelor's degree in International Politics from Georgetown University's School of Foreign Service in the United States and an MA degree in Modern Indian Studies from the University of Gottingen in Germany. His research interests include global developments in artificial intelligence, with particular focus on the risks posed by AI and efforts to build values and norms around future use.