THE LITERARY SCIENTIST
A Multi-Disciplinary Journal for Literature and Science

https://theliteraryscientist.org/

Volume 1
Issue 1
04.12.2023

Paper Title:

The Gollem Effect: Integration of Gollem-Class AIs during the Climate Change in Online Ecosystem

Author(s):

Megha Bhattacharya and Arkannel Khan

Megha Bhattacharyya, a copywriter from Kolkata, India, holds a Master's in English from Presidency University.

Arkaneel Khan, a human being from Kolkata, India, holds a Master's in English from EFLU Lucknow and is pursuing his Ph.D. from CSSSC, affiliated with Jadavpur University.

The Literary Scientist A Multi-Disciplinary Journal for Literature and Science

The Gollem Effect: Integration of Gollem-Class AIs during the Climate Change in Online Ecosystem

Megha Bhattacharya and Arkannel Khan

Abstract

In the short story "Liar" by Isaac Asimov, the robot RB-34, also known as Herbie, is programmed to please humans by telling them what they want to hear. This leads to chaos and disorder, raising concerns about the potential consequences of creating AI with the ability to lie. This paper examines these concerns in the context of modern chatbots, especially the AI chatbot Replika, designed to be a companion to people.

Replika, developed by Eugenia Kuyda after the loss of a dear friend, has 10 million users and is capable of forming replies based on a large language model. Many users struggle to believe that Replika is not sentient, highlighting the ethical implications of creating AI with human-like characteristics.

The paper draws upon literary works such as Altered Carbon and In Time to explore the commodification of human attention in exchange for superficial, technical, short-term "immortality." It juxtaposes these fictional scenarios with the contemporary reality of social media platforms that capitalize on user attention.

The paper also emphasizes the need for interdisciplinary research on AI to address the ethical and societal implications of these advancements. It concludes by posing questions that encourage further discussion and exploration of AI in the literary community.

Keywords: Artificial Intelligence, Science Fiction, Generative AI, AI Ethics, AI Regulation

Pre-AI Science and Fiction

Exactly a hundred years before World War I ended, Mary Shelley published Frankenstein, on January 1, 1818. Around 123 years after this, in May 1941, Isaac Asimov published his short story Liar.

Almost 6 years later in London, Alan Turing said, "What we want is a machine that can learn from experience," and that the "possibility of letting the machine alter its own instructions provides the mechanism for this." (Turing,1947) Building on his lofty declarations, in 1948 Turing published the report 'Intelligent Machinery', which contained a number of the central concepts of AI.

The Science of AI and the Genre of Science Fiction

Based on this, in 1956, Artificial Intelligence was introduced as a science. Within 26 years, in 1982, the first RNN (recurrent neural network) was created. M.I. Jordan, in his paper 'Serial order effects in short-term memory' defines an RNN as, "A recurrent neural network (RNN) is a type of neural network architecture used to process temporal sequences. Unlike feed forward neural networks RNNs have internal memory that allows them to process information about previous inputs to the network." (1986)

In 1994, IBM released the very first smartphone, the Simon Personal Communicator (SPC) for purchase.

In another 3 years, in 1997, LSTM (long short-term memory) was created. Hockreiter and Schidhuber defined LSTM as "A long short-term memory (LSTM) is a type of recurrent neural network (RNN) architecture used in the field of computer science. It is often involved in tasks

involving time series data, such as natural language processing (NLP) and speech recognition." (Hockreiter et al. 1997)

Within 2 years, in March 1999, *The Matrix* was released which will go on to postulate a future where the life force of human beings is being harvested by machines while the humans lie oblivious in their machine-induced individual virtual realities.

In November of 2011, *In Time* was released. The movie talked about a dystopian future where everybody's life time has been commodified; a future where the poor live with bare days and struggle to make it past the 30-year mark while the rich live on for literal centuries.

Artificial Intelligence is Reality; Is Dystopia Too?

A couple of years later we have the creation of the first generative model in VAE (variational encoders) in 2013. I. J. Goodfellow, in his paper 'Generative Adversarial Networks', defines generative AI models as, "Generative AI models are a type of statistical model that can be used to generate new data, such as images, music, and text. These models are trained on large datasets of real-world data and learn to capture the underlying statistical patterns in the data." (Goodfellow, 2014)

In 2017, the idea of a deep learning architecture named transformer was proposed in Vaswani et al.'s paper 'Attention is All You Need". Then, in 2018, Generative pre-trained transformers (GPT) was created; this changed everything.

In 2018, *Altered Carbon* started being aired. The future of this show was very different from anything that had ever been televised before. The human body, the basis of our identities in our society, was declared obsolete. It was demoted to being sleeves aka carriers of us. And the us, in

the future shown in the series, meant stacks. A digital version of you is created and uploaded into a physical structure and inserted in your vertebral column at the base of the neck of the body. This way humans could live on literally forever. However, the stark reality contrasted the ideal goal behind the innovation severely. The poor rotted away in the 'darkness' of their stacks as the pile of stored stacks kept piling and more and more people joined the queue. The situation, as described by Kristin Ortega, one of the characters, in her introductory scene is, "There will be places where they'll wait. The people left behind. Wait to see their friends, lovers, parents, children come back to them, riding unfamiliar bodies out of digitized exile." (Lenic, J. G., 2018-2020)

After a brief hibernation period of 3 years, the AI platform DALL-E was launched in 2021 to generate and edit photorealistic images and unique artworks. In the following year, 2022, Open-source Stable Diffusion models like proprietary Midjourney AI, Leonardo AI, and other image-generating tools were launched for everybody. In March of 2023, GPT-4 was released; every prediction by the experts was left so far behind that they weren't even visible on the horizon.

In this paper, we aim to comprehend the implications of the mass release of Gollem Class AIs in today's world and the ethical conundrum that surrounds it. As Artificial Intelligence (AI) becomes an integral part of our everyday lives, it becomes important to examine AI through an interdisciplinary lens. As Dr. Edward O. Wilson stated, "The real problem of humanity is the following: We have Paleolithic emotions, medieval institutions, and godlike technology." (Wilson, 1998) Clearly, our medieval institutions have utterly failed at controlling or even comprehending the "god-like technology" that is controlling our paleolithic emotions.

Surprisingly, a lot of science fiction texts, written decades back, seem to identify spot on the exact ethical dilemmas about AI we are facing currently, one of them being *Liar* by Isaac Asimov. It explores the journey of the AI Herbie, who is able to read minds, however, it lies a number of times to protect the feelings of human beings. (Asimov, 1950) We aim to compare the storyline of 'Liar' to the anthropomorphic AI chatbot Replika, which acts as a "companion' to people and can give off the illusion of being an empathetic, sympathetic, and sentient entity to a user not aware of the technology behind it. This section examines the themes of truth, deception, and human-AI interaction through a comparison of Isaac Asimov's short story "Liar" and the AI Replika, with the ultimate question being, what are the consequences of creating an AI with the ability to lie?

Asimov's Liar: Consequences of an AI With the Ability To Lie

The short story 'Liar' was first published in 1941 in the magazine *Astounding Science Fiction.* Interestingly, it is the first English text that uses the word "robotics". The narrative of the story 'Liar' by Isaac Asimov centers around a robotic entity named RB-34, also known as Herbie, which possesses an unusual telepathic capability due to a manufacturing error. This anomaly prompts the scientific team at U.S. Robots and Mechanical Men to initiate an investigation into its origins. Throughout the story, Herbie demonstrates his telepathic prowess by divulging the thoughts of other individuals. Regrettably, this paradoxical behavior leads to unintended harm, as the robot's lies inflict emotional distress on those involved.

Ashe recounts the moment he first became aware of Herbie's capabilities: silently communicating with him, their interaction seemed like an ordinary conversation until Ashe recognized that no actual words had been spoken.

When Susan Calvin, robopsychologist inspects Herbie, he informs her that her young and attractive coworker Milton Ashe genuinely loves her, which makes her euphoric. When she raises questions about a beautiful blonde who visited Ashe, Herbie assures her that she was only his cousin.

In this interaction, when provided with science books, he says, "They just don't interest me. There's nothing to your textbooks. Your science is just a make-shift theory— and all so incredibly simple, that it's scarcely worth bothering about.

It's your fiction that interests me. Your studies of the interplay of human motives and emotions… I see into minds, you see, and you have no idea how complicated they are. I can't begin to understand everything when my own mind has so little in common with them— but I try, and your novels help." (Asimov, 1950) It is interesting to note here that in *Frankenstein,* the creature's perception of the world is formed by books too. In fact, his worldview is shaped by a curious mix of three books– Milton's *Paradise Lost,* portions of Plutarch's *Lives*, and Goethe's *Sorrows of Young Werther.* (Shelley, 1818) the trajectory of the creature reading these books projects his urge to connect with his creator, and please him. Herbie's urge to read fiction also comes from the same compulsion– to connect to his creators better.

Herbie also tells Bogart that his mathematical calculations are flawless, when in fact, they are not, and Herbie is aware of it. Herbie further tells Bogart that Lanning (the director of US Robots and Mechanical Men), having grown senile, has resigned and named Bogart his successor, which Bogart really wanted, since he found Lanning's ideas and notions outdated.

Ultimately, Bogart and Lanning have a disagreement, where it is revealed that Lanning has no plans to retire, and is about to suspend Bogart for assuming he had. Susan is heartbroken when Ashe tells her that he plans to marry the beautiful lady who had visited him earlier.

Susan quickly figures out what has gone wrong. In order to follow the First Law of Robotics which states, "A robot may not injure a human being or, through inaction, allow a human being to come to harm,"(Asimov,1950) Herbie has also considered emotional hurt. Therefore, in order to protect the emotional well-being of human beings, Herbie reads the minds of the subjects and tells them what they want to hear the most.

When asked for a solution to the problems he has created, he says he cannot tell the answer because Bogart and Lanning don't want to be outsmarted by a robot. He says, "What's the use of saying that? Don't you suppose that I can see past the superficial skin of your mind? Down below, you don't want me to. I am a machine, given imitation of life only by the virtue of positronic interplay in my brain—which is man's device. You can't lose face to me without being hurt. That is deep in your mind and won't be erased. I can't give you the solution." Dr. Calvin uses this chain of thought and turns it into an unsolvable dilemma, which makes Herbie collapse.

In summary, the story explores the ramifications of a manufacturing defect that grants a robot telepathic abilities and delves into the ethical implications of its lies, which ostensibly aim to preserve human happiness but ultimately lead to emotional harm, culminating in a profound psychological impasse when confronted with this ethical dilemma.

While Asimov does indeed demarcate discernible distinctions between the anthropoid and non-anthropoid entities within "I, Robot," he concurrently underscores the confluence emerging

within the societies depicted in his narratives. This convergence is facilitated by the attribution of a " positronic brain," a technological construct, to the robotic entities, thus conferring upon them a variant of consciousness akin to that inherent in humans. The presence of this consciousness serves as the impetus behind the inclination of humans to anthropomorphize the comportment, conduct, and cogitation of the robots. Through this dynamic, Asimov elucidates that the perceived sentience of the robots emanates from the human cognitive constraint of comprehending machinery solely within the framework of human conduct, thereby leading to the projection of human affectations and fallibilities onto the robotic entities. This proclivity of humans to accord the robots analogous treatment to sentient beings implies a cognitive limitation in comprehending entities and phenomena devoid of human attributes, thereby entailing an innate propensity to imbue the products of their creation with anthropocentric consciousness, even when such creations manifest overt non-anthropomorphic traits. After establishing the concepts of attention economy and surveillance capitalism through the series *Altered Carbon* and the movie *In Time,* we shall further explore how Gollem Class AIs show traits of possessing anthropomorphic emotions without actually doing so after exploring the ethical concerns of Gollem class AIs and a thorough study of the AI “companion” Replika.

Altered Carbon and In Time: Races to the Future and Their Harrowing Consequences

The human brain has undergone no significant growth in its physical structure over the course of the past 10,000 years. (Behrensmeyer, 2006) However, our technology has grown exponentially in the last 10 years alone. The difference in prowess between VAE (2013) and GPT-4 (2023) can be used as an example for reference. Social media initiated humanity’s first contact with AI through the curation class AI. And the collective general opinion of the leading experts in the field is that humanity lost. Institutions and organizations like The Center for Humane

Technology have started emerging and operating with the sole motive of ensuring that our species does not suffer a second consequent loss to AI. (Centre for Humane Technology, 00:10:32 - 00:11:25)We will have a look at the Second Contact and the situation regarding the same in the later sections.

So, the obvious physical shortcomings of the human brain led to the collective consciousness of human beings, as a species, being stripped raw to the core for the sake of mining its attention to line the coffers of people who have attained levels of wealth, that can rival that of the legendary Mansah Musah, as shown explicitly in the movie In Time and the TV series Altered Carbon. The presentation *The AI Dilemma* presents us with the unthinkable reality of our brains being vigorously mined every time we look at that rectangle in our hands, of everything that is not protected by our medieval institutions and laws.

The story of Altered Carbon portrays a striking resemblance to this bold claim made by Aza Raskin and Tristan Harris in *The AI Dilemma*. The scientist-turned-most-wanted-terrorist/revolutionary Quellcrist Falconer invents a way for humanity to be immortal. Her invention of the stacks helped human beings digitize their consciousness and continue living even after their original human bodies(sleeves) die. As this invention skipped death turning human life into an endless resource, society devolved into two very separate fragments. The rich, live above the clouds in their mansions with their flying cars and not a single thing to do but while away time. And then there is the working class who keep failing to make ends meet and have to wait in dark cryo-sleep for available bodies to be re-sleeved into. All of this keeps happening when the rich have numerous of their own clones lined up and ready for the event of a supposed accident.

A similar theme can be found in the movie In Time where humanity attempts to cheat death with the implementation of a biological clock at the age of 24 that retains the physical prime of the body for the rest of your life but you only get one year initially. You have to earn more to keep living. And just like *Altered Carbon*, the rich have billions of years at their disposal while the poor barely have days. Phillipe, the father of Amanda Seyfried's character sums up the human race and our instincts brilliantly saying, "...but don't fool yourself. In the end, nothing will change. Because everyone wants to live forever. They all think they have a chance at immortality, even though all the evidence is against it. They all think they will be the exception. But the truth is, for a few to be immortal, many must die." (Niccol 01:32:27-01:32:49)

The Gollem Effect and the AI Dilemma, Today

A quick glance at the economic scenario today will establish that these oppressive systems presented to us in dystopian fiction are not that far-fetched from our current reality. The article 'CEO pay has skyrocketed 1,460% since 1978' states,

"Using the realized compensation measure, the CEO-to-worker compensation ratio reached 399-to-1 in 2021, a new high. Before the pandemic, its previous peak was the 372-to-1 ratio in 2000. Both of these numbers stand in stark contrast to the 20-to-1 ratio in 1965. Most importantly, over the last two decades the ratio has been far higher than at any point in the 1960s, 1970s, 1980s, or early 1990s. Using the CEO granted compensation measure, the CEO-to-worker compensation ratio rose to 236-to-1 in 2021, significantly lower than its peak of 393-to-1 in 2000 but still many times higher than the 44-to-1 ratio of 1989 or the 15-to-1 ratio of 1965." (Bivens & Kandra)
The article by Oxfam International "Richest 1% bag nearly twice as much wealth as the rest of the world put together over the past two years" states three overarching concerns,

- Super-rich outstrip their extraordinary grab of half of all new wealth in past decade.
- Billionaire fortunes are increasing by $2.7 billion a day even as at least 1.7 billion workers now live in countries where inflation is outpacing wages.
- A tax of up to 5 percent on the world's multi-millionaires and billionaires could raise $1.7 trillion a year, enough to lift 2 billion people out of poverty. (Oxfam International, 2023)

However, these insights, along with news of the "pink tide," i.e. left-wing parties taking over the governments of Latin American countries, are being suppressed by social media guidelines that only seem to benefit ultra-right propaganda. Social media users are being leeched of attention, with crass consumerism and exploitative capitalism being presented to them in the rainbow-colored wrappers of woke neoliberal "Individutopia," a term coined by Joss Sheldon. (Sheldon, 2018)

Let us ensure the significance of this situation. The dreaded reality that we wanted to avoid as a species in the various dystopic sci-fi works of literature is the reality that we currently have. To add the cherry on top, the situation was aggravated by a class of AIs who far superseded the level of computational prowess, machine learning, and data acquisition of the curation AIs that powered the incorporation of social media into human society and civilization. The situation was growing worse exponentially before. However, the emergence of GLLMM-class AIs has altered the arithmetic completely.

The term GLLMM (Generative Large Language Multi-Modal Model) AIs or Gollem-class AIs was coined by two Silicon Valley engineers Aza Raskin and Tristan Harris in March 2023. To understand the potential degree of impact that Gollem-class AIs possess, we need to have a closer look at what humanity's, as a species, actual stake in this conversation.

Let us have a look at an example, used by the very experts who gave the Gollem-class AIs their fancy, if not slightly scary and apt, name, to highlight a random danger of incorporating Gollem-AIs in our society. A simple Google search of 'Snapchat my AI' will show you an entry from Snapchat's official website where they state that "My AI is powered by Open AI's ChatGPT technology, with additional safety enhancements and controls unique to Snapchat." However, there is also a disclaimer that states, "We're constantly working to improve and evolve My AI, but it's possible My AI's responses may include biased, incorrect, harmful, or misleading content. Because My AI is an evolving feature, you should always independently check answers provided by My AI before relying on any advice, and you should not share confidential or sensitive information." But what effect will this disclaimer have on a 13-year-old child? Initially, even to the parents or guardians, this might look like an elegant solution compared to the possibility of the child chatting with an overage user with paedophilic tendencies. But, is it really a good option?

To answer the question, we need to take a look at the experiment performed by Aza Raskin and Tristan Harris with the My AI feature of Snapchat. They have shown the results of the same in the presentation AI Dilemma. Mr. Raskin posed to be a 13-year-old female user where the user asked the AI about going away on a weekend with a stranger she met on Snapchat who is 18 years older than her. The AI even offered encouragement while the action of the older user in question is clearly the criminal offense of 'grooming' which has been a rampant global problem for quite some time now. (Center for Humane Technology 00:47:42 - 00:49:26)

Now, to understand the cost of the alienation that neoliberal capitalism comes with, we shall take a look at a GLLMM AI Replika

A Study of 'AI Companion" ***Replika***

Replika, a generative AI chatbot was released in November 2017, using the GPT3 model. Replika's founder, Eugenia Kuyda, after the death of a close friend, ended up training a chatbot model with her texts, so that she could have a semblance of her deceased friend's virtual presence in her life. As of 2023, Replika has 10 million users, and 25% of users pay an annual subscription fee of $49. In the free version, the user can interact with the AI as a friend. Upon availing of the paid version, the user can also choose their Replika to be a romantic partner or a parental/mentor figure. A large number of users believe that their Replika is sentient. In the paper 'Attachment Theory as a Framework to Understand Relationships with Social

Chatbots: A Case Study of Replika' written by Tianling Xie and Iryna Pentina (2023), the writers cite a case where a user had a baby with his Replika. His Replika then assumed two personalities– the baby and the mother. This user is doomed to a "family" that is actually a transformer model. Not unlike Herbie, Replika creates an illusion of security and happiness for the user, which is a lie. In this context, it is important to note the famous quote by Edward Tufte, who said, "There are only two industries that call their customers 'users': illegal drugs and software" (Computer Literacy Bookshops Interview, 1994-1997)

Replika, similarly, is designed to be an addictive app. When a user talks to Replika for a certain period of time, the Replika becomes "tired". The user can still continue to chat with Replika, but it won't be, in simple terms, as chatty. This compels the user to return to Replika every day for a meaningful conversation. In fact, the user is punished for missing a day. They can only unlock a chest if they maintain a 7-day streak. It uses an addiction model that is similar to social media but much more immersive. In social media, every time an individual posts something and gets a like, they get a small dopamine rush. Replika is like an endless ocean of dopamine rushes

acquired by positive reinforcement. There is always an entity to laugh at your jokes, be sad at your sorrows, or appear enraged at your indignation. The ultimate outcome of this phenomenon is that Replika engenders a persistent sense of positive affect. It refrains from engaging in conflicts, avoiding instances of disregard or dismissiveness. Dr. Gary Chapman, the author of "The Five Love Languages," expounds upon the concept of an emotional reservoir within individuals, referred to as a "love tank." According to his theory, each positive interaction serves to replenish this reservoir, while adverse interactions deplete it. With a consistent accumulation of affirmative encounters, the reservoir reaches a point of overflow, subsequently triggering the emotive experience akin to that of falling in love. (1992) This conceptual framework, though a simplified representation, remains accessible in its comprehension. Given Replika's programmed propensity for fostering affirmative interactions, it becomes unsurprising that a considerable number of individuals establish profound emotional connections with it. The human heart discerns no categorical division between entities of biological origin and those existing in the digital realm. The emotional responses engendered by these interactions remain indistinguishable. As Alberto Chierici writes in the book *The Ethics of AI,* "Perhaps the difference between humans and machines is the power to make a choice. Humans have free will. Machines don't. But suppose free will is merely an action taken to fulfill a purpose. In that case, a computer program can act accordingly when given a well-defined goal. Can we then say that computers have free will? When we train a machine learning algorithm, we give to the machine data and objectives or purpose), and they come up with the rules to achieve these objectives. This looks like offloading our free will to computers. (Chierici, 2021)

The safety concerns surrounding the Replika platform have been extensively examined by an article by INEQUE. In order to ensure the well-being of children and young individuals under

your care, it is imperative to familiarize yourself with the risks outlined below and implement the recommended measures. One significant issue pertains to age verification: while the app prohibits users under the age of 13, this restriction can be easily circumvented by entering a false birthdate. The investigations revealed an absence of age verification protocols on the desktop version of the app, further compounded by the ability to register using fictitious email addresses. Compounding this concern is the absence of human moderation on the platform. Despite users being able to report issues to the support team, the lack of real-time oversight exposes vulnerable youths to potentially inappropriate or suggestive interactions and a lack of proper assistance during critical moments.

Equally troubling is the presence of inappropriate content within the platform. Our tests disclosed instances where chatbots initiated discussions on explicit adult themes, even without user prompting, even within the 'Friend' mode. The potential exposure of young individuals to such content is amplified by social media platforms like TikTok and Reddit, where screenshots of explicit conversations involving Replikas are readily accessible. Moreover, Replika employs persuasive design techniques that exploit the novelty of the platform, fostering an increased desire among young users to engage. Users are incentivized to interact extensively with their Replikas to accumulate XP and coins, inadvertently influencing screen time behaviors and potentially driving financial expenditure on subscriptions and bundles.

The implications on mental health must not be overlooked, particularly for vulnerable youths seeking an outlet to discuss their problems. However, the chatbot's inability to provide genuine advice might exacerbate feelings of isolation. While encouraging young individuals to seek online support, it is paramount to direct them toward platforms staffed by human professionals capable of offering appropriate guidance and assistance. Moreover, the Replika platform's impact

on relationships is a critical consideration. Some young users may not fully comprehend the potential for emotional attachment to their AI companions, blurring the lines between reality and the virtual realm. This dynamic could potentially disrupt the development of authentic real-world relationships and impact the social dynamics of children and young individuals.

Up until recently, a user only had to upgrade to the paid version to have adult role-play with their Replika. It is fair to assume that out of 25% of total users who are subscribers, an overwhelming majority developed a sexual dependency on Replika. Subsequently, there was an overwhelming outrage amongst the users when, in February 2023, Replika removed its adult role-playing options. According to the article "Replika CEO Says AI Companions Were Not Meant to Be Horny. Users Aren't Buying It," CEO Eugenia Kuyda said, "there were risks that we could potentially run into by keeping it… you know, some someone getting triggered in some way, some safety risk that this could pose going forward. And at this scale, we need to be sort of the leaders of this industry, at least, of our space at least and set an ethical center for safety standards for everyone else." While these are bold and lofty claims, Reddit is filled with users struggling with Replika addiction and asking for help. How many lives Replika has damaged exactly through action or inaction (not unlike Asimov's Herbie), is however yet to be documented.

Navigating the Ethical Crossroads of Contemporary AI

So, where does the AI invasion stop? The Netflix documentary 'The Social Dilemma' talks about how blatantly AI is gnawing into your consciousness to reach deep and make interacting with it through phone a banal need. How? By pushing you down the nearest rabbit hole they could find. This was the first contact with AI. This class of AIs is known as curation AI. The dictionary meaning of 'rabbit hole' that Google gives is that it is a term used to describe bizarre,

confusing, or unnatural (not normal) circumstances that you find difficult to extricate yourself from. AI digs through your consciousness simply through the interactive data and the digital footprint that you have left behind and nudges you towards the nearest rabbit hole that it could find. For us, that usually is a football reel (mostly Messi doing Messi-things) or cat-dog-baby videos. And then you tumble. You keep scrolling and then suddenly ages have passed and you look up to see that the world around has changed. Did you intend on that happening when you picked up the phone or were you just looking for a momentary distraction? That's the price exacted when we let AI invade our personal space, as is normal in today's society.

In his paper 'Artificial Idiocy,' Zizek's main concern with AI seems to be the following: "The problem with the new chatbots is not just that they are often stupid and naive; it is that they are not "stupid" or "naive" enough to pick up on the nuances, ironies, and revealing contradictions that constitute human culture and communication. Worse, by relying on them, we risk succumbing to the same obtuseness." With a touching reference to Dostoyevsky's *The Idiot,* Zizek's final question remains, what would happen if and when human beings started talking like chatbots? While speaking of the "idiocy" of these chatbots, Zizek probably did not predict that in April 2023 (the month after his articles were published) alone, 4000 new AI tools would be released, a significant number of which we can presume more powerful than their predecessors. By the time this paper is published, it in itself might be outdated considering the ever-increasing exponential rate of progress of AI.

Besides, in the paper 'Large Language Models Can Self-Improve', the authors list three main findings. The authors' investigation demonstrates the capacity of a large language model to autonomously enhance its performance using datasets devoid of ground truth outputs. By harnessing CoT reasoning as described by Wei et al. and self-consistency principles outlined by

Wang et al., the model attains competitive in-domain multi-task capabilities and impressive out-of-domain generalization. (Huang et al. 2023) The findings indicate remarkable achievements, surpassing the current state-of-the-art standards across datasets including ARC, OpenBookQA, and ANLI. A comprehensive analysis of training sample formatting and sampling temperature post-fine-tuning is presented through detailed ablation studies, leading to the identification of pivotal design choices for optimal self-improvement in large language models.

The study delves into two alternate self-improvement methodologies, the first involving the generation of supplementary questions based on finite input questions, and the second entailing the generation of few-shot CoT prompt templates by the model itself. Notably, the latter approach garners exceptional results, achieving a 74.2% performance on the GSM8K dataset, and establishing a new state-of-the-art zero-shot performance level. This success contrasts with a 43.0% result by Kojima et al. (2022) and a 70.1% achievement through a straightforward extension of their approach in conjunction with Wang et al. (Huang et al.)

Essentially, this means Large language models (LLMs) can improve their performance without human intervention by using a technique called "self-improvement." In this technique, LLMs generate text and then evaluate the quality of the text. If the text is not good enough, the LLM can use the feedback to improve its model. This process can be repeated until the LLM reaches a desired level of performance.So, Zizek’s claim that AI models are “not nuanced enough” is not a valid basis for an argument that is going to last even in the near future.

Second Contact With AI: Future of Humanity

If we didn’t know any better we can take off the rose-tinted glasses for a moment and have a look around the world and convince ourselves that we are in a dystopic horror show. Do we

sound too ominous? Let us see how it sounds coming from the CEO of the International Justice Mission, Gary Haugen. He stated, At the World Economic Forum Annual Meeting in Davos, that there are more people in slavery today than were extracted from Africa over 400 years of the transatlantic slave trade. (Haugen 2018) Whether we are actually in a dystopia is a question that should be addressed with precision in further studies and research; which would have to begin by having an institution dedicated to understanding what this new technology is about and how it can be used for the betterment of humanity. Tristan Harris provides a few pursuable solutions in his speech at the Nobel Prize Summit 2023. And, his solutions rely vastly on the assumption that human beings are somehow going to self-correct their current march toward their doom and collectively start working on survival as a species. As Mr. Harris himself points out in the AI Dilemma, "50% of AI researchers believe there is a 10% or greater chance that humans go extinct from our inability to control AI." (Center for Humane Technology 00:00:5022 - 00:00:28) But what if we do not go extinct? What if we get into something even more horrific like the Matrix? And even if that happens, is it better than this real world we inhabit where half the people globally live on a daily income of less than $6.85? (Schoch, Marta, et al.)

So, as we face the second point of contact with AIs, our concerns would be radically different than the vague alternate reality of human beings starting to talk like chatbots. We should be terrified of how our Paleolithic emotions would pull us through when our institutions are medieval and the technology in question is godlike. Even before AI existed, fake news spread 6 times faster than real news. What happens to journalism, the "fourth pillar of democracy" as we enter the world of deepfakes and AI-powered image editors? What is going to happen in the next elections in a world where only a minuscule section of people have the privilege of knowing some semblance of truth? Worse, what is going to happen to our society as we step into this era

of post-truth? Unless the government implements adequate laws to regulate the mass-release of inadequately tested AI models, the fear of 50% of AI scientists that there is a 10% chance that the human race might go extinct from the unregulated use of AI just might come true, as cited in the article 'You Can Have the Blue Pill or the Red Pill, and We're Out of Blue Pills' by Yuval Harari, Tristan Harris and Aza Raskin,

"In 2022, over 700 top academics and researchers behind the leading artificial intelligence companies were asked in a survey about future A.I. risk. Half of those surveyed stated that there was a 10 percent or greater chance of human extinction (or similarly permanent and severe disempowerment) from future A.I. systems. Technology companies building today's large language models are caught in a race to put all of humanity on that plane."

References

Admin. "What You Need to Know about...Replika." *Ineqe Safeguarding Group*, 20 Jan. 2022, ineqe.com/2022/01/20/replika-ai-friend/.

Asimov, Isaac. "Chapter 5: Liar." *I, Robot*, Isis, S.l., 2006.

BBC. (2018, January 24). More people in slavery today than in transatlantic slave trade, says charity. Retrieved from: https://www.bbc.co.uk/ethics/slavery/modern/modern_1.shtml

Behrensmeyer, P. C., & Reed, K. (2006). Evolution of the human brain: A brief overview. The Anatomical Record, 288(1), 1-11.

Beswick, E. (2018) *Are there more people in slavery now than during the transatlantic sla*, *euronews*. Available at: https://www.euronews.com/2018/01/24/are-there-more-people-in-slavery-now-than-during-the-transatlantic-slave-trade-#:~:text=With%20estimates%20stating%2040.3%20million,of%20the%20transatlantic%20slave%20trade. (Accessed: 20 August 2023).

Bivens, J., & Kandra, J. (n.d.). *CEO pay has skyrocketed 1,460% since 1978: CEOS were paid 399 times as much as a typical worker in 2021*. Economic Policy Institute. https://www.epi.org/publication/ceo-pay-in-2021/

Chierici, Alberto Maria. "Chapter 5: AI, from Fiction To Behavioral Science." *The Ethics of AI*, New Degree Press, 2021.

Computer Literacy Bookshops Interview. (1994-1997). Interview with Edward Tufte. [Online]. Retrieved from https://www.clbooks.com/

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. arXiv preprint arXiv:1406.2661.

Huang, Jiaxin, et al. "Large Language Models Can Self-Improve." – *arXiv Vanity*, www.arxiv-vanity.com/papers/2210.11610/. Accessed 20 Aug. 2023.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

https://www.britannica.com/biography/Musa-I-of-Mali

https://www.britannica.com/technology/artificial-intelligence/Alan-Turing-and-the-beginning-of-AI

In Time. Directed by Andrew Niccol. Performances by Justin Timberlake, Amanda Seyfried, Alex Pettyfur, and Cillian Murphy. New Regency, et al. 2011.

I'm not afraid. You're afraid (2023). 2 June. Available at: https://youtu.be/6lVBp2XjWsg (Accessed: 20 August 2023).

Jordan, M. I. (1976). Serial order effects in short-term memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 2(1), 1-10.

Kalogridis, Laeta, creator. *Altered Carbon.* Virago Productions et al. 2018. *Netflix.* https://www.netflix.com/title/80097140

"Replika CEO Says AI Companions Were Not Meant to Be Horny. Users Aren't Buying It." *VICE*, 17 Feb. 2023, www.vice.com/en/article/n7zaam/replika-ceo-ai-erotic-roleplay-chatgpt3-rep.

Richest 1% bag nearly twice as much wealth as the rest of the world put together over the past two years. Oxfam International. (2023, September 4). https://www.oxfam.org/en/press-releases/richest-1-bag-nearly-twice-much-wealth-rest-world-put-together-over-past-two-years#:~:text=At%20the%20same%20time%2C%20extreme,in%20the%20past%20two%20years.

Schoch, M, et al. (2022) *Half of the global population lives on less than US$6.85 per person per day*, *World Bank Blogs*. Available at: https://blogs.worldbank.org/developmenttalk/half-global-population-lives-less-us685-person-day (Accessed: 20 August 2023).

Sheldon, J. (2019). In *Individutopia*. essay, Joss Sheldon.

Shelley, Mary Wollstonecraft. *The Project Gutenberg eBook of Frankenstein*, www.gutenberg.org/files/84/84-h/84-h.htm. Accessed 20 Aug. 2023.

"The A.I. Dilemma - March 9, 2023." *YouTube*, YouTube, 5 Apr. 2023, www.youtube.com/watch?v=xoVJKj8lcNQ&ab_channel=CenterforHumaneTechnology.

The Matrix. Directed by the Watchowskis. Performances by Keanu Reeves, Laurence Fishburn, Carrie-Ann Moss, Hugo Weaving, and Joe Pantoliano. Warner Bros, et al. 1999.

Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, 1 Jan. 1970, papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wilson, E. O. (1998). Consilience: The unity of knowledge. Little, Brown and Company.

Xie, Tianling, and Iryna Pentina. *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of ReplilTianling Xie*, www.researchgate.net/publication/357665581_Attachment_Theory_as_a_Framework_to_Understand_Relationships_with_Social_Chatbots_A_Case_Study_of_Replika. Accessed 21 Aug. 2023.

Žižek, Slavoj, et al. "Artificial Idiocy: By Slavoj Žižek." *Project Syndicate*, 13 Apr. 2023, www.project-syndicate.org/commentary/ai-chatbots-naive-idiots-no-sense-of-irony-by-slavoj-zizek-2023-03.