Exploring the Risks of General-Purpose AI: The Role of the Brain's Reward Mechanism and Nearsighted Goals in Processes of Decision-Makings

Deivide Garcia da Silva Oliveira[1][0000-0002-5004-1949]

[1] Federal University of Reconcavo of Bahia, Cruz das Almas, BA, 44380, BRA
deividegso@gmail.com

Abstract. The introduction of general purposes AI (such as, but not only, LLMs-large language models AI) in our daily lives draws everyone's attention due to its almost unlimited possibilities. Concerning the risks AI brings to us humans, collectively and individually, we argue that especial attention should be given to the vulnerability of our processes of decision-makings. These processes depend on many things, and we focused on two. First how AI can take advantage of things like our brain's rewards mechanism. Second, we address the question of how the type of goal, if nearsighted or farsighted, is relevant for our rewards mechanism, due to its fast production of dopamine, acting as a weak spot for an easier manipulation of our decision by AI. We conclude with three suggestions for reducing our exposition to AI's manipulation.

Keywords:. Human-AI interaction,

1 Introduction

There are many problems arising after the introduction of general purposes AI (such as, but not only, LLMs-large language models AI) because of its worldwide harmful capabilities, which have been calling everyone's attention since LLM-AI's, like ChatGPT, was released [1, 2]. In order to address part of this problem associated with control over AI, our specific objective is to investigate the potential risks and implications of nearsighted goals and the brain's reward mechanism in the context of general purpose AI systems (AIs that can perform any task requested). We aim to understand how AI can exploit human decision-making processes based on emotional motivations and the dopamine control resulting from the human brain's reward mechanism. By this exploitation, AI can influence emotions, it can shape beliefs, preferences, and even ideologies. Through this investigation, we seek to raise awareness about the potential consequences of AI's ability to change human behaviors, beliefs and decisions, taking another way of thinking about how LLMs pose risks for individuals and societies. At the end of this paper, we also propose a potential solution for this problem.

2 The relationship between the brain's mechanisms and systems of decisions

Before we explain how AI can participate in this problem, let us explain the relationships between decision-making processes, goals, and reward mechanisms. Approaching this first relationship will make more clear the explanation of how AI exploits such relationships. According to the literature, human goals are established according to dual decision-making processes called Systems 1 and 2 [3, 4]. System-1 argues that the processes to reach a decision "are often automatic, fast, and easily affected by emotion" [5, p.1]. Therefore, in many senses, it is more susceptible to errors, precipitations, and profound manipulations. In System-2, on the other side, the processes are slower and more rational, "in which the most important function of System 2 is the successful override of System 1" [5, p.1]. This scientific view on decision-making processes reflects one of the oldest debates in philosophy, introduced by Plato, on the dispute between rationality versus irrationality (often portrayed by emotions), as which one guides our decisions. In many ways, these dual decision-making processes are the exploitable routes and obstacles for external agents to face when attempting to interfere with our decisions.

In addition to these dual-processes of decision-making theories (slow and fast decisions), another thing involved in these processes is the evolutionary fact that making a decision involves both environmental and biological factors, like genetic, evolutionary, and brain mechanisms. In this sense, to make a decision means to deal with these factors while looking for something else, something we generally refer to as an accomplishment of a goal and its reward, whether immediate or not. Ultimately, the reward is a byproduct of accomplishing a goal, at least from the viewpoint of our brain's evolutionary biology. For instance, scoring a three-point in a basketball game is a goal, but it is not the reward in itself. After all, if a player never loses a three-point shot, her reward perception will decline since she has established a new dopamine baseline.

From a brain's biochemical viewpoint, to accomplish a goal, we not only choose between two general routes, fast and slow, respecting the decision-making processes. We also look for rewards usually associated with such decisions, i.e., dopamine. Part of these processes occurs in our brain through the complex interactions of many parts of the limbic system (the limbic system is composed of many structures, like the prefrontal cortex, basal ganglia, the striatum, and the core of it is the nucleus accumbens, which together control the command reward through the release of dopamine). Thus, what we perceive as a reward will impact dopamine release and may vary from time to time, context to context, and person to person. However, average human expectations indicate what is worthy of pursuing as a reward and what is not, like whatever humans see as painful (like burning a hand) and pleasure, which gives the perception of reward (such as scoring points in a game or accomplishing desired goals).

Thus, to choose which goal we will pursue in different aspects of our lives, many things play a role, two of them being the emotions and the perception of

reward produced by realizing a goal. Naturally, the faster we accomplish a goal, the faster we get the reward and release of dopamine, and the stronger the loop stimulus to stay on that track.

Furthermore, the goals we establish for our lives are related to some form of time and energy spent to achieve them, which also have evolutionary roots [6]. For instance, take the case Lisle and Goldhamer gave [7, p.89] about our dopamine levels regarding our experience with food. According to them, we fall into what they call the pleasure trap. The pleasure trap represents one stage, among five others, of dietary behavior, where a person changes whole natural foods for junk foods and then tries to stop eating junk food. During the natural food stage, the dopamine level is normal (the current established baseline, which is not too high or too low), but when she switches to junk food, the dopamine level increases significantly. However, if she tries to return to whole food, the level of dopamine does not go to the previous stage (normal). Instead, dopamine levels will drop below normal. This is why most people feel trapped in junk food and fail to stop eating junk food.

2.1 Nearsighted and farsighted goals in decision-making processes

Unsurprisingly, sitting on a couch while eating some fatty and delicious food, as happens to many of us, rather than finding dopamine in more healthy and sweaty ways, is a faster and more energy-saving way to get our reward system to light up. Sitting on a couch is one way to say that nearsighted goals can be easily attached to faster rewards, especially if we let the immediacy of our reward system, constantly looking for dopamine, make the decisions.

That being said, combining our brain's reward mechanism with System-1 and System-2 of decision processes opens the door for us to talk about decisions from the perspective of nearsighted and farsighted goals. We can graduate our goals between short and long-distance goals or just farsighted and nearsighted goals. We know that nearsighted goals commonly have faster rewards and emotional appeal. Consequently, an easy way to get our reward is by letting the urgency of our dopamine necessity guide our decisions.

However, what about the prefrontal cortex (where most planning and strategic decisions are made)? Does it not interfere with our choice of goals to get us farsighted and rational goals with more significant and long-term rewards? Yes, PFC can change the preferences of goals and rewards; after all, System-2 can override System-1. Nonetheless, PFC can also improve strategies for accomplishing nearsighted goals and get us faster rewards. In other words, PFC can be put at the service of nearsighted goals. The path PFC will go down depends on many items of the mentioned internal and external factors. The limbic system helps to command rewards by releasing dopamine, which is why, in the end, the PFC can help find ways for "optimized action plans for maximizing reward outcomes" [8, p.27]. If our strategies fail to produce reward and expected dopamine, the reward mechanism signals our prefrontal cortex (PFC) to switch strategies and find ways to meet the current dopamine baseline. As we said, with the exact kind of influence, nearsighted goals might end up being the main target of our prefrontal cortex, and

the brain's reward mechanism will be stimulated due to the immediate generation of dopamine neurotransmitters mainly controlled by levels of dopamine [9].

3 Human-AI interaction: brain's reward under AI's influence

Now, we can turn back to human-AI interaction. For the sake of this research, the critical question is, what kind of external factors could change the direction of our choices and even our decision-making processes? Among these factors, cognitive training [5, p.1] and external agencies play an essential role. In other words, other intelligent agents can influence us. For instance, one person can influence another, although this may be one of the most challenging and uncertain things to realize. As a test, try now to change someone's musical preferences, and let us see how it goes. Therefore, the disturbing question is, can external influence (like a human agent) change our personal beliefs, preferences, and even ideologies? It can, although challenging to say the least, especially with subjective matters like music, movies, and ideologies. They are subjective choices linked to our identities, values, and worldviews. So, changing peoples' musical preferences or ideologies is achievable through arguments between humans, although it is extremely hard. Achieving it successfully would require a deep understanding of who we are and our mental buttons as individuals and species. It would also only be possible after gaining our trust (such as through previous good instances of accomplished goals we delivered to AI. Thus, we would demand that such an agent holds some general human values, like being seen as trustworthy, but we would also want this agent to be efficient in solving problems beyond our imagination, to be skilled, and objective. Also, we would demand this agent to be relentless and sharply focused on realizing our desires, *individually speaking*. Such an agent cannot possibly be organic, after all, this kind of unique dedication towards someone's interests and supposedly altruistic behavior is not common in organic agents. However, non-organic agents, like LLMs, precisely promise to fulfill. Literature has shown that inorganic artificial intelligence is the leading candidate because of its building structure made to pursue our desires individually and to understand us (in the broad sense of the word), both as species and individuals [2, 10]. Machines increase their understanding of humans daily to accomplish our goals. Of course, the only problem with this building structure and understandability is the dualistic nature of LLMs, which is simultaneously harmful and helpful. For instance, Stuart Russell even tells us that once we have general functional purpose AIs, these machines will have a "much greater understanding of human psychology, beliefs, and motivations, [and] it should be relatively easy to gradually guide us in directions that increase the degree of satisfaction of the machine's objectives" [11, p.139].

We know that, historically, adult humans do not have their minds easily changed by others for various reasons. However, for the first time in the history of our species, technological innovations made non-organic agents, like AI, possible and highly capable of changing our minds. General purposes AI (able to realize basically any task requested) is advancing fast. Due to its structure being made to pursue each person's goals individually while also being able to understand human psychology collectively, then our behaviors, beliefs and decisions are much more

susceptible to AI than to human agents due to AI's higher precision about who we are and what we want (gradually built). Many scientists and Big Data workers have already expressed the concern that AI can indeed influence our behavior and change our beliefs, even affecting very subjective values, such as music preferences, ideologies, datable partners, and our personal beliefs [12, 13, 14, 15]. LLMs do this by refining information about our brain mechanisms, neurological tricks, historical data of individual choices, social engineering about specific matters, algorithm recommendations, neural networks, deep learning machine systems, emergent capacities of AI, and more.

In a recent paper published by many authors, among them Google and OpenAI engineers (the company that developed ChatGPT), the authors warn us about LLMs' emergent capacities for deceiving people and changing people's beliefs. Accordingly, "the model is effective at shaping people's beliefs, in dialogue and other settings (e.g. social media posts), even towards untrue beliefs." [2, p.4]. In a few words, AI cannot only help us to achieve our goals, whether nearsighted or farsighted goals, but it can also influence us based on our previous decisions and emotions (tracking and quantifying each choice we make), i.e., seeking and finding the fastest route to dopamine release.

AI operates mainly through nearsighted and emotionally motivated goals, which is extremely dangerous for all individuals and countries. Underdeveloped countries, like Brazil, with high levels of government corruption, poverty, and judicial system fragility, become easier targets for these technologies [16]. For instance, AI can quickly push specific narratives during long periods towards developing and employing "sequential plans that involve multiple steps, unfolding over long time horizons" [2, p.4]. AIs can do this in order to slowly and unnoticeably change our beliefs bit by bit and small decision by small decision. Adomavicious and collaborators [12] have shown how this is possible even with non-general purposes AI, like simple recommendation algorithms about music preferences. Imagine what advanced general purpose AIs can do! Also, it is noteworthy to recall one of the reasons our emotions can control our decisions with the help of AI. The primary reason is that we do not like to be displeased, which is partially rooted in the neurological mechanism of the brain's reward [17]. The researchers C.O'Connor and J.Weatherall [18] have shown that dissonant information with our beliefs is one reason for the spread of misinformation, even before the internet. We usually avoid cognitive dissonance coming from the collected information. We like to look for bias-supporting information, i.e., we "tend to seek out attitude-consistent information and avoid attitude-challenging information." [17, p.3]. Whatever pleases a person is also what triggers her dopamine levels. Obviously, a general purpose AI will learn this as quickly as it happens and fine-tune its actions, like personal recommendations, to meet whatever pleases the end users. For instance, during COVID-19, Brazilians were exposed to untrue beliefs and misinformations, and part of the reason they spread out misinformations is that these beliefs met with a significant part of the population's nearsighted goals (like going to parties or Malls, attaching to political groups) and also their system of beliefs (like the simple fact the cope with this new reality is something nobody wanted to

do, and that liberty means to do whatever a person wants). At least half of the population felt rewarded and joyed when found on social media influencers, the medical community, and political and religious leaders blessing them to do what they already wanted to do. Naturally, the encounter between these leaders and half of the population on social media platforms was not a coincidence but mostly the work of AI systems.

Nevertheless, how do we avoid having brain mechanisms, like the reward mechanism, exposed to AI systems so much? The first thing to notice is that we must be aware that we humans, collectively and individually, are more vulnerable to advanced AI systems than we can imagine. So, no matter what we do or how we think, there will always be significant human vulnerability to these technologies. After all, we are dealing with an agent who can understand our psychology better than ourselves [11]. The second thing to notice is that this vulnerability comes in degrees, according to a myriad of items, and it is also in constant change due to ongoing human-AI interaction. For instance, it depends on what model and advancement of AI is being used, the goals pursued by the human agent using AI, the conflicts arising between various goals of multiple agents, the size of the database available, how good the training and learning program applied is, the decision made according to the results obtained, how vital the subject matter is for us, how tired we will be to make decisions, the kind of choices engineers take interpreting the data collected, and so on. The third and last thing to notice is that whatever choice we make to keep AI systems in check, whether by better laws, better algorithms, better technology education, or better policies, the aim of keeping AI systems in check is somewhat similar to the effort we daily employ to keep our own irrationality, emotional beliefs, and wild desires in check. We must be employ critical thinking, constantly questioning and checking the feedbacks, and reducing our dependency of AI. To reject this duty in practice means delegating our decisions to a third party (AI) while choosing intellectual laziness, although remaining accountable and affected by the consequences. Therefore, human-AI interaction will not be humanly favored without our active and constant human control over AI, and the risks henceforward emerged will not go away regardless of how good the legal, political, or scientific informed support we end up having.

4 Conclusion

In conclusion, our research sheds light on the potential risks posed by general purpose AI systems by examining the role of nearsighted goals and the brain's reward mechanism on human decision-making processes, making the susceptibility of individuals and societies to AI-driven manipulation. The ability of AI to shape beliefs, preferences, and ideologies raises concerns about long-term consequences for social and personal well-being and freedom.[1]

---

[1] Declaration of conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest whether financial or personal nature.

Bibliography

1. Kosinski, M.: Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083 (2023)
2. Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N.: Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324 (2023)
3. Stanovich, K.E., West, R.F.: On the relative independence of thinking biases and cognitive ability. Journal of personality and social psychology 94, 672 (2008)
4. Evans, J.S.B., Stanovich, K.E.: Dual-process theories of higher cognition: Advancing the debate. Perspectives on psychological science 8, 223-241 (2013)
5. Xu, P., Wu, D., Chen, Y., Wang, Z., Xiao, W.: The effect of response inhibition training on risky decision-making task performance. Frontiers in Psychology 11, 1806 (2020)
6. Godfrey-Smith, P.: Darwinian populations and natural selection. Oxford University Press (2009)
7. Lisle, D.J., Goldhamer, A.: The Pleasure Trap: Mastering the Force that Undermines Health & Happiness. Publisher. Book Publishing Company (2003)
8. Sesack, S.R., Grace, A.A.: Cortico-basal ganglia reward network: microcircuitry. Neuropsychopharmacology 35, 27-47 (2010)
9. Lewis, R.G., Florio, E., Punzo, D., Borrelli, E.: The Brain's Reward System in Health and Disease. Adv Exp Med Biol 1344, 57-69 (2021)10.1007/978-3-030-81147-1_4
10. Nowotny, H.: In AI we trust: power, illusion and control of predictive algorithms. Polity, Cambridge, UK (2021)
11. Russell, S.: Human compatible: Artificial intelligence and the problem of control. Penguin (2019)
12. Adomavicius, G., Bockstedt, J.C., Curley, S.P., Zhang, J.: Do recommender systems manipulate consumer preferences? A study of anchoring effects. Information Systems Research 24, 956-975 (2013)
13. Hawking, S., Tegmark, M., Russell, S., Wilczek, F.: Transcending complacency on superintelligent machines. Huffington Post 19, (2014)
14. Adomavicius, G., Bockstedt, J.C., Curley, S.P., Zhang, J.: Effects of online recommendations on consumers' willingness to pay. Information Systems Research 29, 84-102 (2017)
15. Paul, K.: Letter signed by Elon Musk demanding AI research pause sparks controversy. The Guardian, London. (2023)
16. Ribeiro, P.V.: Brasileiros Ganham Frações de Centavos para Melhorar sua Inteligência Artificial. The Intercept, https://www.intercept.com.br/2023/06/19/brasileiros-ganham-fracoes-de-centavos-para-melhorar-sua-inteligencia-artificial/ (2023-Jun-19)
17. Metzger, M.J., Hartsell, E.H., Flanagin, A.J.: Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. Communication Research 47, 3-28 (2020)
18. O'Connor, C., Weatherall, J.O.: The misinformation age. Yale University Press (2019)