Dopamine, Mimicry, and Value Alignment: Artificial Intelligence and Addiction

Atticus D. Rex

Abstract

Experts have voiced ethical concerns of Artificial Intelligence (AI) for decades. This paper provides a framework and a review of how AI can be addictive by mimicking neuropleasurable stimulus for humans. It argues that this problem is exacerbated when the values of the humans interacting with AI and the values of the AI are poorly aligned. The paper examines the role of the dopaminergic reward system in powerful behavior reinforcement, what economic gains and externalities this causes, and how this may disproportionately affect certain subsets of the population. This paper also proposes a model to illustrate how, when considering factors of time, effort, and expected reward, the magnitude of the reward is not deterministic of human action. Lastly, the paper examines ethical frameworks and contemporary theory governing the use of AI to produce large-scale addiction, big-nudging and general manipulation.

A report presented as the Capstone Project for the
Philosophy, Politics, and Economics (PPE) program
at Virginia Tech.

“Any sufficiently advanced technology is indistinguishable from magic."

Arthur C. Clarke

“The best way to keep a prisoner from escaping is to make sure he never knows he's in prison.”

Fyodor Dostoevsky

“Properly speaking the artist, the writer, and the scientist should be moved by such an irresistible impulse to create that, even if they were not paid for their work, they would be willing to pay to get the chance to do it.”

Norbert Wiener

I. Introduction

Due to an exponential increase in computational speed and parallelism in the 21st century, powerful new methods for information processing have become feasible [1]. In addition, according to a 2019 report by the World Economic Forum, humans will have generated approximately 44-64 zettabytes of data in 2020 [2]. To contextualize this number, humans created more data from 2010 to 2011 than every single human ever up to that date [2]. And with this explosion of information, arises a need to process, organize and learn from it. Artificial Intelligence (AI) will be defined henceforth as the ability of computers to perform intelligent tasks in place of humans with minimal human intervention [1]. Intelligence will be defined as the ability to learn from data and then accurately predict the outcomes of its environment and, subsequently, produce actions that increase the probability of preferable changes to its environment. Another important distinction for the scope of this report is the difference between AI and Machine Learning (ML). Machine Learning is a type of AI but, specifically, is a process in which machines adapt their own behavior or predictions to match a set of data [3]. Not all AIs are adaptive, but the most powerful AIs can learn and recognize unseen patterns in vast amounts of data.

Today, one of the most successful types of machine learning is a technique called “Deep Learning Neural Networks” [4]. The variants of neural network architecture loosely mimic the structure of human neurons to propagate information from some input to some output, like the famous large language models do [5]. There are two salient points about this type of machine learning: first, it can be rigorously proven that given enough neurons and enough data, deep neural networks (DNNs) can model *any* input to *any* output, regardless of how complex the relationship is, meaning their long-term potential to learn from data is theoretically infinite [6].

Second, DNNs are essentially black boxes that are exceedingly difficult to recover any semantic insight from the patterns they discover, which makes them cumbersome to evaluate for bias and specific learning [6]. For example, to know what OpenAI's GPT-3 has "learned", one would have to sift through 175 billion intertwined parameters [5]. With this in mind, AI will be transformative in solving a plethora of problems that have plagued humans for centuries. Whether it is the protein folding algorithm, AlphaZero, image and video recognition, fraud detection, or energy and environmental solutions, AI will improve human civilization in untold ways [7]–[10]. However, experts in the field have been quick to voice concerns about the potential of AI to do humans harm since even the early days of computing [1]. Problems such as militarization, mass manipulation, surveillance, social sorting, unmonitored or unsafe implementations, nonconsensual testing, data and privacy breaches, and exacerbated inequality are all present today in the conversation of AI ethics [1].

This paper will focus on two main subproblems of AI: drive-mimicry and value alignment. Drive-mimicry is the act of AI mimicking phenomena that humans, by nature, find pleasurable. Directly, this phenomenon produces experiences that imitate social interaction, or sexual content. However, AI will also learn to synthesize new types of pleasurable things under the umbrella of psychoactive substances such as food, drink, and new drugs. Value alignment is the idea that, as AI systems get more and more complex, the goals of humans become more and more difficult to explain to a computer [11]. An example of this might be a social media algorithm that has been programmed to maximize a human's time on the platform. One could argue that if a human wants to spend more time on the app, then his goals are aligned with that of the AI. But what if the AI maximizes the person's time on the app by providing more and more addictive, or polarizing material, is it really in the person's best interest to stay on the app?

Humans have struggled with right and wrong since the beginning of recorded history and it is naive to think that they will be able to implement an AI that works perfectly in humans' best interest. This paper will delve into the economic and ethical implications that such problematic AIs pose when they are distributed at scale, drawing from different disciplines like modern psychology to substantiate the claims. To keep a concise scope, this paper will mainly be focused on the United States, but these problems are specific to humans, not necessarily Americans.

II. Thesis

AI drive mimicry will target the habit-forming, pleasure center of the human brain. This mechanism involves the release of the neurotransmitter dopamine, which is the same chemical responsible for reinforcing behavior like eating, drinking, socializing, caring for offspring, and reproducing [12]. Targeting this process reduces humans' ability to make decisions that their brain prefers in a purely cognitive sense by stimulating the "hot" motivation enough to silence the "cold" motivational system. This effect is more pronounced in susceptible subgroups such as children, low-income earners, the mentally ill, and the elderly. The value-alignment problem will further exacerbate this phenomenon. Because of the inherent potency of dopaminergic learning, the mass eroding of economic rationality within a population as a result of harmful AI, and the disproportionate effect that AI-based manipulation poses on vulnerable parts of a population, the United States economy will be less productive, and stray from a socially desirable outcome with AI used in this way. Further, it is immoral to implement AI in this way because of the psychological distress it causes, the moral opportunity cost for those educated enough to implement this type of AI and its aforementioned impact on economic prosperity.

III. Psychological Approach

In a 1971 experiment, researchers first noticed that damage to dopamine pathways caused a reduction in feeding and drinking in rats [13]. A working hypothesis began that dopamine simply affected the motor functional pathways of rats which prevented them from taking physically doing these activities [13]. However, by the late ’70s, after many trials of dopamine agonists and antagonists, it became quite apparent that dopamine was responsible for *motivation* and had no effect on motor impairment [12], [14], [15]. Since this foundational research, the neurotransmitter has been causally linked to pleasurable behavior and reinforcement in mammals—humans are no exception [16]. These pleasurable behaviors include eating and drinking, sexual experience, positive social interaction (experiences like praise, social bonds, and empathy), achievement, and caring for young [12], [14]. These types of behavior are quintessential for the survival and reproduction of humans, and, over the past few million years, the dopaminergic reward system has developed into a powerful mechanism to motivate certain behaviors [17].

More problematically, however, humans have discovered ways to release this neurotransmitter artificially in the form of dopaminergic drugs such as stimulants (cocaine, nicotine, etc.), amphetamines, and various opiates [18]. An important distinction to draw is that not all drugs release dopamine to cause neuropleasurable experiences; a host of drugs mimic the release of serotonin to produce pleasurable experiences such as MDMA, psilocybin, and LSD [19]. However, modern research indicates that these serotonergic drugs do not reliably cause habit-forming behavior, though research on the true neurobiological processes is still ongoing [20]. Conversely, there are thousands of studies to show the replicable habit-forming effects of the aforementioned dopaminergic drugs [21].

Addiction is formerly characterized in the DSM-5 as being “physically and mentally dependent on a particular substance, and unable to stop taking it without incurring adverse effects.” [22]. In this paper, it will be adapted to: “physically or mentally dependent on a particular substance or activity, characterized by a decreased ability to stop taking the substance, which can incur adverse effects”. The adjustment comes from the recent assertion that not all addictions are physical, nor are all addictions caused by substances.

In recent years, humans have discovered that the artificial release of dopamine comes not only as the result of external drugs but also as the mimicry of these primary survival-based neuropleasurable experiences. And this is where technology finally collides with the reward center of the human brain. Technology has aided humans in mimicking phenomena like social interaction, sexual experience, and achievement in varying degrees over human history. However, the advent of digital computing, and subsequently the internet, have accelerated this process to hyper speed. Digital video games first came into existence in the 1970s, and strived to be as exciting and pleasurable as possible—and they largely realized this pursuit, with a myriad of research indicating clinical addiction to many popular video games [23]. It has been colloquially stated that the first industry to adopt new technology is almost always the pornography industry. Indeed, the porn industry standardized the use of Super 8 film and VHS tapes, and in 2001 there were approximately 21,000 online porn sites in some of the internet’s nascent years [24]. Pornography sites were also some of the first major adopters of virtual reality in the early 2010s [25]. Likewise, the first social network was created in 1997, just 26 years ago, the first domino in a powerful attempt at mimicking and intensifying social interaction [26].

A major hurdle to the scientific study of indirect dopaminergic behavior is the inability to directly measure dopamine release in healthy human subjects in real-time. This causes a disconnect in quantifying *how* addictive these phenomena are on a neurobiological level. However, new technologies are at the forefront of changing this. It will soon become feasible directly compare social media with drugs like cocaine [27].

With AI recently arriving at the scene, so will it be a major player in mimicking social interaction, sexual experience, and entertainment. Indeed, it already has in sites with predictive content like TikTok, Instagram, Snapchat, and YouTube, where AI learns from heaps of user data to craft the perfect suggestions, engagements, and rewards [28]. The dawn of interactive AI has just begun—soon, humans will flock to AI to produce interactive sexual or emotional dialogue, explicit or entertaining photos and videos, and even physical intimacy. Based on the level of addiction that pornography and social media currently pose, AI stands to significantly increase it. Imagine how influential an AI could be if it can read and remember every single publicly available psychology paper in a matter of hours (which it can and has, with OpenAI's GPT-4 just on the horizon at the time of this paper) [5].

Not only has an excess of dopaminergic stimulus been strongly linked to anhedonia (inability to feel pleasure) in mammals, it has also been strongly shown to negate the most fundamental motivations, such as the motivation for life itself [29], [30]. Mammals have been shown to choose dopaminergic experiences over and over until they die of starvation [30]. This is a strong indication that polarizing stimulus to the dopaminergic reward center of the human brain supersedes rational, healthy desires. This should cause quite a bit of concern as the world pours its time and resources into these uncharted waters, with potent new forms of addiction lurking just beneath the surface.

IV. Economic Rationality Approach

Models are essential to economics as simpler “cartoon-versions” of complex phenomena. These help economists understand and make predictions about an infinitely complex world. One of the most fundamental models in economics is the simplification of humans to a “Homo Economicus” [31]. This is a concept that assumes economic actors will be rational and self-interested in all situations. It states that in all scenarios, humans will act to maximize their utility. The term “rationality” in economics simply means that humans make decisions that are consistent with their preferences to maximize their well-being [31]. And for preferences to be considered rational, they merely have to be complete and transitive. This is a rather broad brush and many economists have been quick to point out that there appear to be two distinct types of preferences. Several frameworks describe this. The “Dual System Model” of cognition describes that there are two motivation systems that not only process information differently but also change with age [32]. There has also been the differentiation between “Hot” Executive Function (EF) and “Cold” EF, where Hot EF is responsible for processing exciting, neuropleasurable, emotional stimuli while Cold EF is responsible for purely cognitive processes [33]. Economist Daniel Kahneman, the recipient of the 2002 Nobel Prize in economics, has described these two systems as “system 1” and “system 2”, respectively [34].

Economists have also worked toward a theory of rational addiction—by the laws of completeness and transitivity alone, addiction is perfectly rational [35]. However, scholars in the field have been quick to illustrate the rapidly changing preferences depending on exposure to addictive behavior and instability in long-standing preferences. This necessitates a higher level of refinement than a theoretical vector of preferences that can completely rearrange itself at a given moment based on one external stimulus. As a model, this does not provide enough insight into *why* this rearrangement happens and how to predict it in the future, though this is the result.

The artificial and excessive release of dopamine described in the previous section illustrates how humans may still make rational decisions but rely far more heavily on the hot system as opposed to the cold, which often leads to less desirable outcomes. In the example used previously, I claim it is not in any mammal's best interest to inject cocaine until it starves to death. This is because the cold system, or purely cognitive system, would reject this idea almost immediately. However, this phenomenon has been replicated over and over again in behavioral studies of mammals. Likewise, it is not in the best interest of humans to use heroin—the drug strongly degrades physical health, happiness, and almost all measures of well-being and functionality [36]. However, hundreds of thousands of Americans wake up every day and negate the cold cognition of their brain to use heroin. And they do so for the sole reason that it so strongly stimulates the habit-forming part of the human brain to the point where it causes physical dependencies and essentially completely negates the ability of the cold system to regulate behavior. And to a similar degree, humans often know by way of reason that it is probably not in their best interest to waste hours of one's day on extraneous entertainment, social media, or pornography sites—and yet millions do so each day.

Because of the novelty and complexity of AI and its place in human life, it is difficult to reliably design experiments for animals that test things like AI-accelerated social media and interactive or AI-generated pornography. Thus, proper measurement of addiction and dopamine release in *humans* is crucial to providing strong evidence for the regulation of such technologies. However, fMRI technology and new biotechnology are showing huge promise in differentiating and gauging the activity of these two mechanisms of motivation.

V. A Reward-Effort-Adjusted Theory of Planned Behavior

With the emerging threat of AI targeting the habit-forming, pleasure center of the human brain, comes a need to assess human behavior on a more refined level than “hot” or “cold”. As described, AI-accelerated media can produce human behavior that negates the motivations of the purely cognitive side of the brain, as governed by the Homo Economicus standard economic model. Thus, it is important to operationalize some new utility function that humans try to maximize. In particular, human behavior is not motivated by just the *reward* of something. It is also constrained by the *effort* it takes to achieve the reward. For example, one might perceive the following situations to be extremely rewarding: seeing the northern lights in a five-star hotel, attending a world-class university, owning a successful startup, and so on. While all of these might be extremely rewarding, they require an immense amount of effort, time, and careful planning to achieve. So what humans tend to do in reality is maximize the *ratio* of expected reward to the effort required to achieve it.

Thus, I propose the following model to illustrate the ideas in this paper. Consider the assumptions that all humans attempt to maximize the following adjusted reward function, *F*:

$$F(\text{effort}, \text{time}) = \frac{\text{Expected Reward}}{\text{effort} \cdot \text{time}}$$

Reiterating the above, humans want to maximize the expected reward while minimizing the effort and time required to achieve the said reward. However, it is desirable to implement this mathematically, so the following function will be used:

$$F(\epsilon, t) = \frac{1 + \beta \cdot P(\epsilon, t)}{e^{\epsilon \cdot t}}$$

In this case, ϵ is the intensity of the effort expended, β is the possible reward to be gained from the effort, and *t* is the amount of time expended on the task. $P(\epsilon,\ t)$ is the probability of receiving the reward, β, given the effort, ϵ, and the time, *t*. Hence, $\beta\ \cdot\ P(\epsilon,\ t)$ is the expected reward based on the magnitude of the reward and the probability of receiving the reward. Notice that when ϵ and *t* are both zero, i.e. when zero time or effort is expended, the function yields 1. This should denote a neutral state in which the agent has neither exerted time nor energy and has also not received any expected reward.

In general, it can be assumed that increasing the time and effort to receive a reward increases the probability of receiving it. For example, increasing the intensity and duration at which a student studies for a test increases the probability of doing well on the test. Thus, an approximation to the probability function will be used in this paper as a hyperbolic tangent function. This simulates the law of diminishing returns. The difference between studying zero hours to one hour will likely increase the probability of a good grade quite drastically, but the difference between studying for ten hours to eleven hours probably will not make quite as good of a difference. Thus, the final equation becomes:

$$F(\epsilon, t) = \frac{1 + \beta \cdot \tanh(\alpha \cdot (\epsilon \cdot t))}{e^{\epsilon \cdot t}}$$

In this final function, the parameter, α , represents the rate at which increasing time and effort increases the probability of receiving the reward, β. With this model, a number of observations can be made. Consider the following case where $\beta = 1$ and $\alpha = 3$:

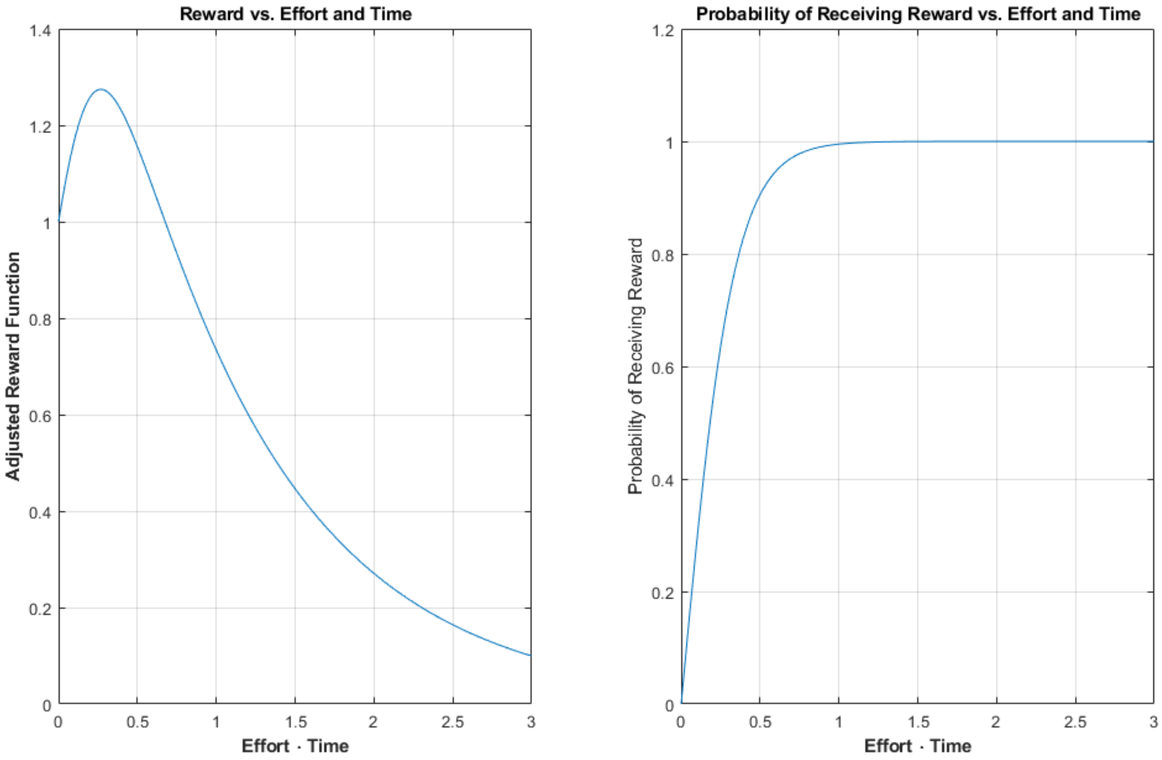

Figure 1: (Left) Plot of Adjusted Reward Function vs. product of effort and time. (Right) Plot of probability of receiving reward vs. product of effort and time.

In the above plot, the value α is quite high. Thus, increasing time and effort greatly increase the probability of receiving the reward. An example of this might be sitting down and watching a movie or scrolling through social media—something that takes almost no effort to achieve the desired reward. Let us consider a different scenario. Now let us consider a higher reward, but a lower α, say $\beta = 5$ and $\alpha = 0.3$:

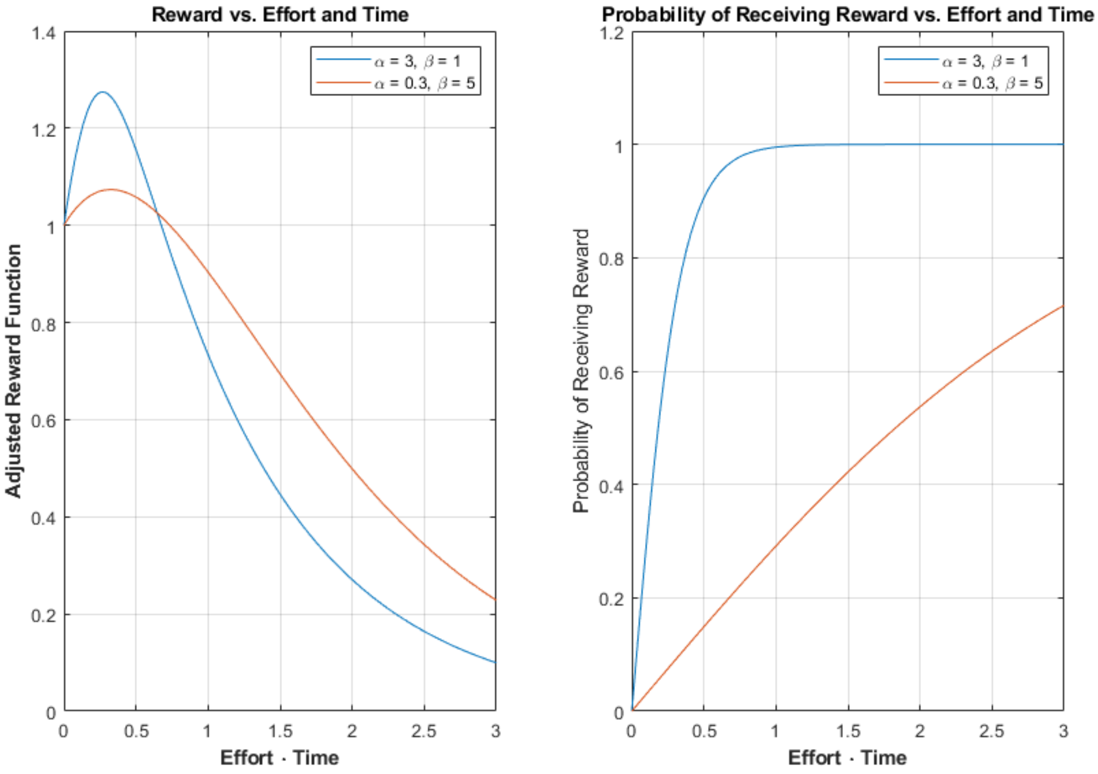

Figure 2: (Left) Plot of Adjusted Reward Function vs. product of effort and time for different levels of alpha and beta. (Right) Plot of probability of receiving reward vs. product of effort and time for different levels of alpha and beta.

Figure 2 illustrates that, even when the reward is multiple times as large, when we adjust for effort and time to receive the reward, the lower reward wins out. And this phenomenon is present in human behavior quite often. It is the reason humans choose to eat junk food instead of cooking healthy meals or spend hours on social media instead of exercising. And this subconscious maximization that has kept humans alive for so many millennia makes drive-mimicry AI with poor value alignment so dangerous.

Some economists argue that humans make decisions following binary decision trees [37]. The idea of rationality is about an agent's list of preferable outcomes in order of favorability. While this is true in a more abstract sense, in practice, humans seem to think more on a bifurcation tree rather than a vector magically constructed in one's head. To give some introduction to this idea, consider the following example of one who is experiencing hunger:

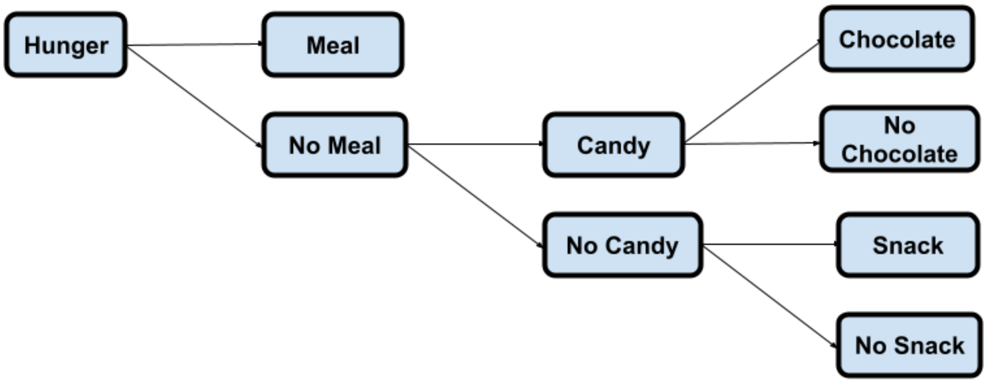

Figure 3: A diagram of the binary decision tree that one might make upon searching for a solution to hunger.

For each outcome in the diagram in Figure 3, the adjusted reward function can be evaluated. While making a full meal might be the best route to sate hunger in a healthy way, the perceived effort and time needed to make a said meal are large—and eating a meal of cooked vegetables might not release as much dopamine as eating a chocolate bar. Hence, thinking through all of these outcomes and evaluating the adjusted reward function at all of them, we can see how an agent might choose chocolate because of its low time, effort, and high dopaminergic reward.

Equating this to microeconomic theory, many of the decisions that humans make are in some way a reflection of this model. The true value in this economic model is synonymous with the number of dopamine molecules that get released between human synapses on a given day. Indeed, almost everything humans choose to do on a daily basis is on some pathway to a feeling of dopamine—and much of this revolves around money because money is the most versatile key to releasing dopamine. A jetski might be fun for a day but money can unlock an infinite number of dopaminergic activities. This model is largely a measure of the hot system—the cold system still has the ability to override it given it is below a certain threshold. However, AI will learn this threshold where humans choose short-term pleasure over long-term by lowering the effort required to achieve it and increasing the neuropleasurability of the experience. This will usher in a new era where instead of producing products to match consumer demand, firms now have the ability to use AI to manipulate consumer demand to match the products.

VI. Economic Gains and Externalities

There are a number of immediate economic externalities that will come as a result of deploying harmful AI on a population. Conversely, there will emerge significant new markets as a result of this new technology advancing at such a rapid rate. Organizations have already started adopting LLMs into their infrastructure to make tasks feel more human-like [38]. From ChatGPT to Microsoft's Bing implementation of GPT-4, new technology is already spawning a huge new market of natural language interaction which will boost productivity [39]. This creates a large new market of untapped value that will likely transform how humans interact with technology. This said the burden of value alignment is always painfully present. Consider a shopping site whose sole purpose is to sell merchandise—this site can now deploy a powerful AI that learns from a user's movements throughout the site to augment the information and media in order to maximize the time a user spends on the app. And it can coordinate all of this data from all users to learn which strategies are effective and which are not—all in real-time on millions of data points. This kind of consumer manipulation is present in many applications and has contributed to the economic success of platforms like TikTok, YouTube, Facebook, Instagram, and many others [40]. At current rates, the average American will spend 5.7 continuous *years* of adult life on social media [26]. Hence, there is enormous economic potential in this vast sink of global consumer attention.

Further, AI will also be revolutionary in the food and drug industries [41]. AI has already demonstrated its might in its ability to solve complex organic molecular interactions, specifically DeepMind's AlphaZero which solved protein folding for the first time—a feat that had prior to eluded humans [8]. This technology will only become more powerful and commonplace within industry to develop new compounds for medicinal use—but also for recreational use. The cannabis industry is still in its nascent stages, but already organizations are finding ways to produce new and selective compounds from the plant [42]. The opiate Fentanyl was synthesized to ease the economic burden on heroin addicts—instead, it has become the lead killer in a domestic opioid epidemic [43]. AI will have a significant hand in altering and formulating new food, drink, and drugs to sell to humans. It is no surprise that the best-selling food, drink, and recreational drugs are also the most dopaminergic [44]. And explicitly or implicitly, AI will be key in producing even more intensely neuropleasurable substances.

Moving to the effect that more pronounced AI-accelerated addiction has on a population, addiction has been linked to a host of health problems, increasing the risk of heart disease, stroke, cancer, and mental disorders [45]. While the aforementioned new markets will increase value, the negative effects of these markets will put an increased strain on healthcare throughout the world. A 1996 report estimated that alcoholism alone costs society $99 billion annually [46].

Further, the success rate for treating addiction today is concerningly low—there is no treatment that is even close to 100% reliable, and even the strongest drug-addiction treatments like methadone are only effective on a small minority of patients [47]. Therefore, the addiction that AI-accelerated drive mimicry stands to cause does not have a known solution—the study of addiction is ongoing and incomplete, with the efficacy of many treatments hotly debated [46].

Further, mass manipulation and value misalignment will lower the individual productivity of society as a whole. Because of the anhedonia and lack of motivation to do mundane tasks that are present in excess dopaminergic stimulus, humans will work less. However, it is difficult to make conclusive claims on whether or not productivity will change based on the increased efficiency that new technology will pose. Humans today work less and produce far more than their ancestors did a mere century ago [48]. But the effects of AI-accelerated addiction are not to be overlooked—indeed, in 2022 Americans spent on average 15,000 years of human life on TikTok alone *each day* [49]. With magnitudes of engagement like this, macroeconomic outcomes are crucial to consider when regulating these hugely powerful AI models with spurious value alignment.

VII. Ethical Implications

The implementation of Artificial Intelligence at scale in new and efficient ways requires intense schooling and study of the discipline. Take the famous new language models like ChatGPT, GPT-4, Bard, and so on, for example. Developing these systems to their current level of sophistication and outreach requires hundreds of experts at the top of their fields in many disciplines such as mathematics, statistics, parallel computing, IT, neuroscience, experimental design, and so on [50]. Because of the inherent biological reinforcement that AI-accelerated drive-mimicry poses, these systems have the potential to be highly demanded, much like any other dopaminergic material such as pornography, drugs, foods like sugar, sex, and social interaction are highly demanded. And indeed, organizations are already forging new roads to implement this type of AI. A website called Pirate Diffusion is already mixing generative AIs to create explicit user-generated content [51].

The potential to manipulate humans using this technology is large. Philosophers such as Kant, Mill, Rousseau, Hobbes, and even Aristotle, have all vehemently underlined the immorality of manipulation for a variety of reasons. Kant argued that it is a moral discretion to use humans as a means to some end, as a violation of the Categorical Imperative [52]. In contrast, John Stuart Mill theorized that manipulation voids an individual's ability to make free and autonomous choices, as a necessary precondition for a moral society [53]. Rousseau dealt with violations of a social contract, Hobbes sought to demonstrate an undermining of social trust, and Aristotle wrote on the difficulties to achieve human virtue under the guise of manipulation [54]–[56]. For brevity within the scope of this paper, manipulation will be deemed *immoral*.

Moreover, a host of modern thinkers have put significant research into the ethical dilemmas that AI poses. A body of literature exists on the governmental use of AI for widespread propaganda and disinformation, social control campaigns, surveillance, facial recognition, sentiment analysis, social sorting, and even improved interrogation techniques [57]–[61]. There have been discussions on how to assign moral and ethical status to nonhuman agents within an infosphere of agents working together [58]. However, there is very little literature about AI drive-mimicry, which appears to be an emerging problem in the discipline [1].

Now, it is important to consider some ways in which AI-accelerated drive mimicry can be used to manipulate that may not meet the eye. At the onset, whether intentional or not, it will cause powerful new forms of addiction. And with addiction, comes a host of other unintended risks that humans face. First, privacy: when engaging with addictive AIs, humans will likely be required to share private information to continue receiving the neuropleasurable content. This information can then be used in a variety of harmful ways to track, expose, or further manipulate the user. Second, dehumanization: as mentioned in the psychological approach, too much dopaminergic behavior has been strongly linked to anhedonia and a lack of response to more mundane stimulation and motivation [62]. Deploying addictive AI at scale will cause such phenomena in swaths of people all over the world, causing a measurable decline in the quality of human life.

Because of its direct interaction with the reward mechanics of the brain, drive-mimicry technology is in extremely high demand. Software developers, researchers, psychologists, engineers, and many other highly educated people around the world are paid large sums of money to implement these systems—and this trend seems to be increasing [63]. However, because of the inherent immorality of the manipulation and addiction that these systems cause, it is an unethical use of education and skillset to develop this kind of AI. Furthermore, AI stands to revolutionize previously impenetrable areas of human life—in disciplines such as medicine, education, therapy, clean energy, agriculture, poverty, and environmental preservation. The ability of people to use their skills in AI to work towards a transcendent good is enormous. Thus, a sort of moral opportunity cost is formed—talented, educated humans are turning down opportunities to work on projects that would almost certainly improve human life, to get paid large sums of money to develop potentially extremely harmful AI. It may be economical on an individual level to choose this line of work, but it is a moral failure to work toward developing something that can lower the quality of millions of human lives in the way that strong addiction can. This is neither a socially nor morally desirable outcome.

VIII. Equity Considerations

One last important consideration of the ethical and economic concerns associated with AI-accelerated drive mimicry is how it will likely affect different demographics of the population. A host of literature shows that in children, economic hardship and poverty are strongly correlated with the prevalence of mental illness [64]. Further, minorities and elder populations are likewise disproportionately susceptible to mental disorders and distress [65]. Social media and technology addiction can exacerbate this trend. Children use social media sites at a much higher rate than adults do [66]. And further, children also experience the anhedonia associated with excess dopamine stimulus from social media—a 2011 report first coined the term "Facebook depression" for a depressive state brought on by excessive use of social media [66].

A significant precondition for safe AI interaction is the ability to set boundaries to limit exposure to dopaminergic stimuli. Children tend to be less effective than adults at setting healthy boundaries with neuropleasurable experiences, as many studies on impulse control demonstrate [67]. Thus, minimizing child exposure to AI-accelerated social media sites, entertainment, and even explicit sexual content is vastly important to diminish harm to future generations. Children tend to be especially susceptible to technology addiction and, without proper limits and guidance, this can lead to adverse mental health consequences [68]. Further, those unable to access or afford proper education and care are also disproportionately affected by technology addiction. A study on technology addiction in Italy found that lower-income students not only were more susceptible but suffered worse mental health outcomes as a result of technology addiction [69]. Lastly, elderly populations are also a statistically higher target for technology misinformation schemes and other such scams. Elderly populations (over the age of 65) were much more likely to be influenced by misinformation during the COVID-19 pandemic about vaccination information than younger generations were [70].

Hence, it goes to infer that AI manipulation will likely be much more pronounced in these more susceptible generations. This realization is necessary for allocating resources to prevent negative outcomes from affecting the well-being and economic prosperity of those who stand to be affected in the most pronounced way.

IX. Conclusion

Humans are outpacing previous predictions for milestones on the way to Artificial General Intelligence on a weekly basis. OpenAI's GPT-4 is showing "Sparks of General Intelligence" according to a paper published by Microsoft—a feat roughly estimated to take place in 2030 [39]. Not only is the sophistication and breadth of Artificial Intelligence increasing at a breakneck pace, but accessibility is also rapidly growing. This accessibility is to both consumers and developers all over the world. Hundreds of startups have begun implementing large language models, generative image models, and predictive behavioral models. Chatbots, photo, and video generation, suggestive algorithms, and a host of other technologies have fundamentally altered how humans interact with technology. The evidence is already clear that AI will reshape and accelerate study in many, many disciplines like medicine, engineering, agriculture, environmental preservation, and so on. This paper examines the effect of AI that mimics primary neuropleasurable experiences of humans like social interaction, sexual experience, food and drink, and even dopaminergic drugs. This type of AI stands to significantly amplify the already addictive nature of entertainment, social media, pornography, processed food, and addictive substances. Because these types of experiences target the dopaminergic reinforcement system of the human brain, they stand to cause the negative symptoms associated with too much dopaminergic stimulus such as a lack of pleasure and motivation to increase well-being. This report illustrates the failures of the fundamental model when an agent's rational preferences do not reflect her actions and proposes a new framework to assess behavior. The economic impacts of this phenomenon are significant—while this will open new markets, this will also cause significant stress on healthcare systems and the negative effects of addiction are pronounced on human health, well-being, and productivity. Because of the value misalignment, where human values are not aligned with the profit, attention, or interaction-maximizing values of AI, the potential for manipulation and privacy breaches stemming from this reinforced behavior are imminent. In addition, because of how intensely AI stands to make society better and the disproportionate effect on vulnerable demographics, it is immoral to implement an intelligent system to manipulate humans like this. This report examines a fundamental conflict between a huge demand for interactive, pleasurable, accelerated AI systems and a desire for human health and well-being.

X. References

[1] T. Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Minds Mach.*, vol. 30, no. 1, pp. 99–120, Mar. 2020, doi: 10.1007/s11023-020-09517-8.

[2] “How much data is generated each day?,” *World Economic Forum*, Apr. 17, 2019. https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/ (accessed Apr. 11, 2023).

[3] Z.-H. Zhou, *Machine Learning*. Springer Nature, 2021.

[4] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[5] OpenAI, “GPT-4 Technical Report.” arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.08774.

[6] S. Sonoda and N. Murata, “Neural network with unbounded activation functions is universal approximator,” *Appl. Comput. Harmon. Anal.*, vol. 43, no. 2, pp. 233–268, Sep. 2017, doi: 10.1016/j.acha.2015.12.005.

[7] S. Zhong *et al.*, “Machine Learning: New Ideas and Tools in Environmental Science and Engineering,” *Environ. Sci. Technol.*, vol. 55, no. 19, pp. 12741–12754, Oct. 2021, doi: 10.1021/acs.est.1c01339.

[8] H. Zhang and T. Yu, “AlphaZero,” in *Deep Reinforcement Learning: Fundamentals, Research and Applications*, H. Dong, Z. Ding, and S. Zhang, Eds., Singapore: Springer, 2020, pp. 391–415. doi: 10.1007/978-981-15-4095-0_15.

[9] S. Hijazi, R. Kumar, and C. Rowen, “Using Convolutional Neural Networks for Image Recognition”.

[10] Y. H. Liu, “Feature Extraction and Image Recognition with Convolutional Neural Networks,” *J. Phys. Conf. Ser.*, vol. 1087, no. 6, p. 062032, Sep. 2018, doi: 10.1088/1742-6596/1087/6/062032.

[11] I. Gabriel, “Artificial Intelligence, Values, and Alignment,” *Minds Mach.*, vol. 30, no. 3, pp. 411–437, Sep. 2020, doi: 10.1007/s11023-020-09539-2.

[12] R. A. Wise, “Dopamine, learning and motivation,” *Nat. Rev. Neurosci.*, vol. 5, no. 6, Art. no. 6, Jun. 2004, doi: 10.1038/nrn1406.

[13] U. Ungerstedt, “Adipsia and Aphagia after 6-Hydroxydopamine Induced Degeneration of the Nigro-striatal Dopamine System,” *Acta Physiol. Scand.*, vol. 82, no. S367, pp. 95–122, 1971, doi: 10.1111/j.1365-201X.1971.tb11001.x.

[14] “Dopamine and desire,” *https://www.apa.org*. https://www.apa.org/monitor/mar05/dopamine (accessed Apr. 12, 2023).

[15] “Blockade of cocaine reinforcement in rats with the dopamine receptor blocker pimozide, but not with the noradrenergic blockers phentolamine or phenoxybenzamine.” https://psycnet.apa.org/record/1979-05529-001?doi=1 (accessed Apr. 12, 2023).

[16] W. Schultz, “Behavioral dopamine signals,” *Trends Neurosci.*, vol. 30, no. 5, pp. 203–210, May 2007, doi: 10.1016/j.tins.2007.03.007.

[17] F. H. Previc, *The dopaminergic mind in human evolution and history*. in The dopaminergic mind in human evolution and history. New York, NY, US: Cambridge University Press, 2009, pp. ix, 214. doi: 10.1017/CBO9780511581366.

[18] R. Maldonado, A. Saiardi, O. Valverde, T. A. Samad, B. P. Roques, and E. Borrelli, “Absence of opiate rewarding effects in mice lacking dopamine D2 receptors,” *Nature*, vol. 388, no. 6642, Art. no. 6642, Aug. 1997, doi: 10.1038/41567.

[19] S. H. Ahmed, M. Lenoir, and K. Guillem, “Neurobiology of addiction versus drug use driven by lack of choice,” *Curr. Opin. Neurobiol.*, vol. 23, no. 4, pp. 581–587, Aug. 2013, doi: 10.1016/j.conb.2013.01.028.

[20] R. J. DiLeone, J. R. Taylor, and M. R. Picciotto, “The drive to eat: comparisons and distinctions between mechanisms of food reward and drug addiction,” *Nat. Neurosci.*, vol. 15, no. 10, Art. no. 10, Oct. 2012, doi: 10.1038/nn.3202.

[21] G. Di Chiara, “The role of dopamine in drug abuse viewed from the perspective of its role in motivation,” *Drug Alcohol Depend.*, vol. 38, no. 2, pp. 95–137, May 1995, doi: 10.1016/0376-8716(95)01118-I.

[22] American Psychiatric Association and American Psychiatric Association, Eds., *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th ed. Washington, D.C: American Psychiatric Association, 2013.

[23] A. M. Weinstein, “Computer and Video Game Addiction—A Comparison between Game Users and Non-Game Users,” *Am. J. Drug Alcohol Abuse*, vol. 36, no. 5, pp. 268–276, Aug. 2010, doi: 10.3109/00952990.2010.491879.

[24] P. Bakker and S. Taalas, “The Irresistible Rise of Porn: The Untold Story of a Global Industry,” 2007.

[25] M. Orel, “Escaping reality and touring for pleasure: the future of virtual reality pornography,” *Porn Stud.*, vol. 7, no. 4, pp. 449–453, Oct. 2020, doi: 10.1080/23268743.2020.1777895.

[26] “The rise of social media - Our World in Data.” https://ourworldindata.org/rise-of-social-media (accessed Apr. 12, 2023).

[27] B. Le Foll, A. A. Wilson, A. Graff, I. Boileau, and P. Di Ciano, “Recent methods for measuring dopamine D3 receptor occupancy in vivo: importance for drug development,” *Front. Pharmacol.*, vol. 5, 2014, Accessed: Apr. 12, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphar.2014.00161

[28] S. Ben, "How TikTok Reads Your Mind - The New York Times." https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html (accessed Apr. 11, 2023).

[29] I. of M. (US) C. on O. in D. A. Research, *Drug Abuse Research in Historical Perspective*. National Academies Press (US), 1996. Accessed: Apr. 12, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK232965/

[30] W. Roy, *Neural Mechanisms on the Reinforcing Agent of Cocaine*. National Institute on Drug Abuse, 1976.

[31] D. A. Urbina and A. Ruiz-Villaverde, "A Critical Review of Homo Economicus from Five Approaches," *Am. J. Econ. Sociol.*, vol. 78, no. 1, pp. 63–93, 2019, doi: 10.1111/ajes.12258.

[32] L. Steinberg, "A dual systems model of adolescent risk-taking," *Dev. Psychobiol.*, vol. 52, no. 3, pp. 216–224, 2010, doi: 10.1002/dev.20445.

[33] J. P. Roiser and B. J. Sahakian, "Hot and cold cognition in depression," *CNS Spectr.*, vol. 18, no. 3, pp. 139–149, Jun. 2013, doi: 10.1017/S1092852913000072.

[34] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[35] G. S. Becker and K. M. Murphy, "A Theory of Rational Addiction," *J. Polit. Econ.*, vol. 96, no. 4, pp. 675–700, 1988.

[36] J. J. Platt and C. Labate, *Heroin addiction: Theory, research and treatment*. in Heroin addiction: Theory, research and treatment. Oxford, England: John Wiley & Sons, 1976, pp. xvi, 417.

[37] G. A. Akerlof and J. L. Yellen, "Rational Models of Irrational Behavior," *Am. Econ. Rev.*, vol. 77, no. 2, pp. 137–142, May 1987.

[38] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.

[39] S. Bubeck *et al.*, "Sparks of Artificial General Intelligence: Early experiments with GPT-4." arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.12712.

[40] E. Kalampokis, E. Tambouris, and K. Tarabanis, "Understanding the predictive power of social media," *Internet Res.*, vol. 23, no. 5, pp. 544–559, Jan. 2013, doi: 10.1108/IntR-06-2012-0114.

[41] Z. Chen, X. Liu, W. Hogan, E. Shenkman, and J. Bian, "Applications of artificial intelligence in drug development using real-world data," *Drug Discov. Today*, vol. 26, no. 5, pp. 1256–1264, May 2021, doi: 10.1016/j.drudis.2020.12.013.

[42] D. Bolognini and R. Ross, “Medical cannabis vs. synthetic cannabinoids: What does the future hold?,” *Clin. Pharmacol. Ther.*, vol. 97, no. 6, pp. 568–570, 2015, doi: 10.1002/cpt.107.

[43] L. Beletsky and C. S. Davis, “Today’s fentanyl crisis: Prohibition’s Iron Law, revisited,” *Int. J. Drug Policy*, vol. 46, pp. 156–159, Aug. 2017, doi: 10.1016/j.drugpo.2017.05.050.

[44] N. M. Avena, P. Rada, and B. G. Hoebel, “Evidence for sugar addiction: Behavioral and neurochemical effects of intermittent, excessive sugar intake,” *Neurosci. Biobehav. Rev.*, vol. 32, no. 1, pp. 20–39, Jan. 2008, doi: 10.1016/j.neubiorev.2007.04.019.

[45] “Addiction and Health,” *National Institute on Drug Abuse*, --. https://nida.nih.gov/publications/drugs-brains-behavior-science-addiction/addiction-health (accessed May 02, 2023).

[46] A. T. McLellan *et al.*, “Evaluating the Effectiveness of Addiction Treatments: Reasonable Expectations, Appropriate Comparisons,” *Milbank Q.*, vol. 74, no. 1, pp. 51–85, 1996, doi: 10.2307/3350433.

[47] A. T. McLellan, I. O. Arndt, D. S. Metzger, G. E. Woody, and C. P. O’Brlen, “The Effects of Psychosocial Services in Substance Abuse Treatment,” *Addict. Nurs. Netw.*, vol. 5, no. 2, pp. 38–47, Jan. 1993, doi: 10.3109/10884609309149701.

[48] R. Premraj, M. Shepperd, B. Kitchenham, and P. Forselius, “An empirical analysis of software productivity over time,” in *11th IEEE International Software Metrics Symposium (METRICS’05)*, Sep. 2005, p. 10 pp. – 37. doi: 10.1109/METRICS.2005.8.

[49] “U.S. time per day on Netflix, TikTok, YouTube 2024,” *Statista*. https://www.statista.com/statistics/1359403/us-time-spent-per-day-netflix-tiktok-youtube/ (accessed May 02, 2023).

[50] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[51] “Free ai image maker - free ai images generator - the best ai image maker - home,” *Pirate Diffusion*. https://www.piratediffusion.com/ (accessed May 02, 2023).

[52] H. J. Paton, *The Categorical Imperative: A Study in Kant’s Moral Philosophy*. University of Pennsylvania Press, 1971.

[53] J. S. Mill, *The Basic Writings of John Stuart Mill: On Liberty, The Subjection of Women and Utilitarianism*. Random House Publishing Group, 2010.

[54] J.-J. Rousseau and V. Gourevitch, *Rousseau: The Social Contract and Other Later Political Writings*. Cambridge University Press, 2018.

[55] T. Hobbes, *Hobbes’s Leviathan*. Рипол Классик, 1967.

[56] G. Remer, “Rhetoric, Emotional Manipulation, and Political MoralityThe Modern Relevance of Cicero vis-à-vis Aristotle,” *Rhetorica*, vol. 31, no. 4, pp. 402–443, Nov. 2013, doi: 10.1525/RH.2013.31.4.402.

[57] G. Lazer, “CHINA INTERNET REPORT,”

[58] S. Engelmann, M. Chen, F. Fischer, C.-Y. Kao, and J. Grossklags, “Clear Sanctions, Vague Rewards: How China’s Social Credit System Currently Defines ‘Good’ and ‘Bad’ Behavior,” Jan. 2019, pp. 69–78. doi: 10.1145/3287560.3287585.

[59] D. Helbing, *Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution*. Springer, 2018.

[60] J. P. Holdren, A. Bruce, E. Felter, T. Lyons, and M. Garris, “PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE”, [Online]. Available: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

[61] A. McAllister, “Stranger than Science Fiction: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention Against Torture,” *Minn. LAW Rev.*.

[62] M. T. Treadway and D. H. Zald, “Reconsidering anhedonia in depression: Lessons from translational neuroscience,” *Neurosci. Biobehav. Rev.*, vol. 35, no. 3, pp. 537–555, Jan. 2011, doi: 10.1016/j.neubiorev.2010.06.006.

[63] “Salary: Machine Learning Researcher (Mar, 2023) US,” *ZipRecruiter*. https://www.ziprecruiter.com/Salaries/Machine-Learning-Researcher-Salary (accessed May 02, 2023).

[64] R. A. Samaan, “The influences of race, ethnicity, and poverty on the mental health of children,” *J. Health Care Poor Underserved*, vol. 11, no. 1, pp. 100–110, Feb. 2000, doi: 10.1353/hpu.2010.0557.

[65] Office of the Surgeon General (US), Center for Mental Health Services (US), and National Institute of Mental Health (US), *Mental Health: Culture, Race, and Ethnicity: A Supplement to Mental Health: A Report of the Surgeon General*. in Publications and Reports of the Surgeon General. Rockville (MD): Substance Abuse and Mental Health Services Administration (US), 2001. Accessed: May 02, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK44243/

[66] G. S. O’Keeffe, K. Clarke-Pearson, and Council on Communications and Media, “The Impact of Social Media on Children, Adolescents, and Families,” *Pediatrics*, vol. 127, no. 4, pp. 800–804, Apr. 2011, doi: 10.1542/peds.2011-0054.

[67] V. I. Douglas, “Stop, look and listen: The problem of sustained attention and impulse control in hyperactive and normal children,” *Can. J. Behav. Sci. Rev. Can. Sci. Comport.*, vol. 4, pp. 259–282, 1972, doi: 10.1037/h0082313.

[68] D. J. Kuss and M. D. Griffiths, “Online gaming addiction in children and adolescents: A review of empirical research,” *J. Behav. Addict.*, vol. 1, no. 1, pp. 3–22, Mar. 2012, doi: 10.1556/jba.1.2012.1.1.

[69] C. Remondi *et al.*, “Insecure Attachment and Technology Addiction Among Young Adults: The Mediating Role of Impulsivity, Alexithymia, and General Psychological Distress,” *Cyberpsychology Behav. Soc. Netw.*, vol. 23, no. 11, pp. 761–767, Nov. 2020, doi: 10.1089/cyber.2019.0747.

[70] S. C. Chia, F. Lu, and Y. Sun, “Tracking the Influence of Misinformation on Elderly People’s Perceptions and Intention to Accept COVID-19 Vaccines,” *Health Commun.*, vol. 38, no. 5, pp. 855–865, Apr. 2023, doi: 10.1080/10410236.2021.1980251.