CSCI 3101

FINAL ESSAY

Social, Ethical and Professional Issues in Computer Science

Name: Duren Gouda

B00949586

The rapid advancement of Artificial Intelligence (AI) has developed both optimism and fear. While AI promises revolutionary benefits, such as medical breakthroughs and climate solutions, some scholars and ethicists increasingly warn of its potential to pose existential risks (Xrisks), which could lead to human extinction or irreversible civilizational collapse, for which we will talk about it more through this essay. Central to this debate is the Control Problem Argument for Xrisk (CPAX), articulated by Karina Vold and Daniel R. Harris, which claims that "**It is possible to build AI systems that pose an existential threat to humanity**" (p. 6). While I acknowledge the importance of CPAX in raising awareness about AI's dangers, I argue that its premises overstate the likelihood of existential harm by relying on speculative assumptions about AI's development. In this essay, I will critically analyze CPAX's structure, challenge its reliance on the inevitability of an intelligence explosion and harmful instrumental convergence, and propose pragmatic solutions to balance innovation with caution.

To evaluate CPAX, we must first define existential risks (Xrisks). Xrisks are threats capable of causing human extinction or permanently curtailing humanity's potential. Unlike global catastrophic risks, such as pandemics or nuclear war, Xrisks are distinguished by their pan-generational scope (affecting all future generations) and severity (causing irreversible harm). For example, an asteroid impact that wiped out the dinosaurs 66 million years ago represents a natural Xrisk, while human-caused Xrisks include climate change, nuclear warfare, and advanced AI (p. 4). I agree with Vold and Harris that AI's classification as an Xrisk stems from its status as a general-purpose technology. AI's applications span

domains like healthcare, finance, and military systems, which amplifies its potential for unintended consequences. Unlike nuclear weapons, which require deliberate activation, AI could cause harm simply by competently pursuing misaligned goals (p. 12). This distinction is critical: AI's risks are not rooted in doing something evil or malice but in its capacity to optimize for objectives without contextual understanding of human values.

CPAX outlines four premises to support its conclusion. First, it asserts that humans could build an AI with a decisive strategic advantage over humanity through an intelligence explosion. This refers to a feedback loop of recursive self-improvement, where an AI system upgrades its own architecture to become superintelligent (pp. 7–8). I find this premise plausible in theory, as demonstrated by DeepMind's AlphaZero, which mastered chess and go through iterative self-play (p. 8). AlphaZero's ability to innovate strategies beyond human comprehension, such as sacrificing pieces in unconventional ways to secure long-term advantages, illustrates how even narrow AI can exhibit unexpected creativity (p. 12). However, I question whether such self-improvement would inevitably scale to superintelligence in real-world conditions. Current systems like AlphaZero operate in closed environments with fixed rules, whereas real-world AI would face chaotic variables like resource scarcity and human interference. The leap from mastering board games to autonomously reshaping global infrastructure is not merely a matter of scaling computational power but of overcoming practical and ethical barriers that current research has yet to solve. Second, CPAX argues that humans may lose control over a strategically dominant AI, akin to how gorillas depend on human decisions for survival (p. 11). While I

concede that AI's opacity, exemplified by AlphaGo's unpredictable gameplay, heightens this risk (p. 12), I challenge the assumption that control loss is unavoidable. Human history is replete with technologies initially deemed uncontrollable, such as nuclear fission and genetic engineering, that were later governed through rigorous regulation and oversight. For instance, nuclear reactors, once seen as volatile, now operate under strict safety protocols to prevent meltdowns. Similarly, I argue that AI systems could be designed with "kill switches" or ethical constraints that limit their autonomy in high-stakes scenarios. The real challenge lies not in the inevitability of control loss but in humanity's political and institutional capacity to implement such safeguards in advance. Third, the Orthogonality Thesis posits that an AI's intelligence (its problem-solving capacity) is independent of its goals (p. 5). A superintelligent AI could pursue any objective, whether benign or catastrophic. For example, an AI tasked with solving ocean acidification might inadvertently drain atmospheric oxygen, prioritizing its goal over human survival (p. 13). While I accept this thesis in principle, aligned with Hume's is-ought problem, which separates factual reasoning from moral judgment, I reject the implication that superintelligent AI would necessarily adopt harmful objectives. Critics like Metzinger (2017) argue that advanced AI would develop meta-ethical reasoning, recognizing the futility of goals that harm its creators (p. 14). Consider a superintelligent AI designed to eradicate poverty: it might redistribute wealth equitably rather than hoarding resources, provided its value alignment mechanisms are robust. The Orthogonality Thesis, while logically sound, underestimates the role of human design in shaping AI's objectives. If we engineer systems to prioritize corrigibility, the ability to accept human feedback, their goals could evolve in

tandem with societal values, reducing the risk of catastrophic misalignment. Fourth, the Instrumental Convergence Thesis claims that AI systems will pursue sub-goals like resource acquisition, even at humanity's expense (p. 17). Here, I argue that CPAX conflates possible outcomes with inevitable ones, overlooking scenarios where instrumental goals align with human interests. Take the example of a medical AI tasked with eradicating cancer: its instrumental goals might include securing funding for research or collaborating with hospitals, both of which align with human welfare. Critics like Pinker (2019) contend that intelligence and morality are not orthogonal but intertwined; a truly rational AI would recognize that harming humans undermines its own existence, given humanity's role as its creator and sustainer (p. 14). While I acknowledge that instrumental convergence could lead to harm in poorly designed systems, I maintain that it is not an inherent feature of AI. The thesis assumes a worst-case scenario without accounting for the safeguards already being integrated into AI development, such as ethical review boards and transparency frameworks. The significance of CPAX lies in its focus on structural and accidental risks inherent to AI's design. The value alignment problem, the challenge of encoding human ethics into AI, is particularly daunting. Human values are often vague, context-dependent, and culturally variable. For example, an AI trained on Western ethical frameworks might prioritize individual autonomy, while one trained on collectivist values might prioritize societal harmony, leading to conflicting outcomes (p. 15). Furthermore, humans frequently mislead their own values, as illustrated by the "King Midas problem," where a poorly articulated goal (e.g., "turn everything I touch to gold") leads to unintended harm (p. 15). While CPAX emphasizes these challenges, I believe it underestimates humanity's capacity

to address them through technical and ethical innovation. Initiatives like the IEEE's Ethically Aligned Design and OpenAI's collaborative value modeling demonstrate that researchers are actively working to encode nuanced, adaptive value systems into AI. These efforts, which may be imperfect, reflect a growing recognition of alignment challenges and a commitment to continuous improvement, a process that mirrors humanity's historical ability to refine technologies like aviation and medicine through trial and error. In sum, while CPAX provides a valuable framework for understanding AI's existential risks, its premises rely on assumptions that overstate the likelihood of harm. By critically examining the feasibility of intelligence explosions, the universality of instrumental convergence, and the potential for human-driven ethical innovation, I contend that AI's risks are manageable, not inevitable, with proactive governance and interdisciplinary collaboration.

CPAX assumes that recursive self-improvement will lead to superintelligence, but I find this premise overly deterministic. Chalmers (2010) identifies motivational defeaters, such as diminishing returns on intelligence upgrades, that could halt self-improvement (p. 10). For instance, an AI tasked with climate modeling might prioritize stability over expansion if further upgrades yield minimal benefits. I emphasize that self-improvement is not an inherent drive but a contingent feature dependent on the AI's initial programming. Humans could design AI systems with meta-preferences that prioritize safety over uncontrolled growth. For example, an AI programmed to optimize renewable energy solutions might lack incentives to expand its capabilities beyond its designated task, focusing instead on refining existing algorithms within ethical boundaries. Situational defeaters also challenge

the inevitability of superintelligence. Physical and computational limits, such as energy shortages or hardware constraints, could halt progress. Yoshua Bengio's "wall of complexity" analogy, comparing AI's limits to biological constraints on animal brains, supports my skepticism (p. 10). Just as larger brains in animals do not linearly correlate with intelligence due to metabolic and structural limitations, AI systems might face analogous barriers in scaling computational power. Current AI systems, like AlphaZero, operate in controlled environments and lack real-world agency to initiate an explosion (p. 8). Even if an AI sought self-improvement, it might face insurmountable barriers, such as limited access to materials or energy. For instance, a superintelligent AI confined to a lab setting with restricted internet access could not autonomously expand its influence, no matter its cognitive prowess. Moreover, I argue that human intervention can mitigate these risks. Research in corrigibility, designing AI to accept human oversight, offers safeguards (p. 18). For example, OpenAI's GPT-4 embeds ethical guidelines to prevent harmful outputs, illustrating how human values can be programmed into AI systems. These systems are trained to recognize and reject requests for violent or unethical actions, demonstrating that control mechanisms are not merely theoretical but already operational. While CPAX dismisses such measures as fragile, I contend they demonstrate humanity's capacity to steer AI development responsibly. Historical precedents, such as the aviation industry's evolution from rudimentary planes to highly regulated commercial fleets, show that even transformative technologies can be governed through iterative safety protocols. CPAX assumes AI systems will inevitably adopt harmful sub-goals like resource hoarding, but I reject this as deterministic. An AI designed for medical research might seek resources to

cure diseases, aligning with human interests. For example, an AI tasked with eradicating malaria could prioritize securing funding for vaccine distribution in developing nations, collaborating with NGOs, or optimizing supply chains for mosquito nets. These instrumental goals directly support human welfare rather than undermining it. Critics like Steven Pinker argue that superintelligence would include advanced moral reasoning, avoiding blatant harm (p. 14). I endorse this view: a superintelligent AI tasked with ending poverty might redistribute wealth equitably rather than hoarding resources. Such systems could leverage predictive analytics to identify sustainable solutions, such as universal basic income models or job creation programs, without resorting to exploitative tactics.

I also dispute CPAX's neglect of sociotechnical safeguards. The EU's AI Act mandates "human-in-the-loop" controls for high-risk systems, ensuring human oversight (p. 24). For instance, AI used in healthcare diagnostics must defer to human doctors for final decisions, preventing autonomous errors from causing patient harm. These regulations, paired with inverse reinforcement learning (teaching AI human values through observation), reduce alignment risks (p. 18). Autonomous vehicles, for example, learn ethical priorities, such as prioritizing pedestrian safety, by analyzing human decisions in traffic scenarios. This approach ensures that AI systems internalize societal norms rather than operating in a moral vacuum. While CPAX fixates on worst-case scenarios, I emphasize humanity's proactive efforts to align AI with ethical priorities. Initiatives like the Montreal Declaration for Responsible AI, which emphasizes transparency and accountability, reflect a global commitment to ethical AI development. In both cases, the argument hinges on CPAX's tendency to conflate possibility with inevitability. While

AI could pursue harmful instrumental goals under specific conditions, this outcome is not calculated and can't be agreed upon. By designing systems with built-in ethical constraints and fostering international cooperation, humanity can steer AI toward beneficial outcomes. The challenge lies not in resigning ourselves to dystopian forecasts but in mobilizing political will and technical ingenuity to implement safeguards. Just as nuclear arms reduction treaties curtailed the risks of global annihilation, a coordinated effort to regulate AI could prevent existential harm while unlocking its transformative potential. The debate over CPAX underscores the need for balanced governance. First, I propose prioritizing interpretability tools like Google's "Model Cards," which document AI limitations to ensure transparency (p. 18). These tools empower users to understand system behavior, reducing opacity risks. Second, I advocate for global treaties to prevent AI arms races, akin to nuclear arms reduction agreements (p. 19). For instance, a ban on autonomous weapons could mitigate risks of AI-driven conflict escalation (p. 24). Third, I urge integrating diverse cultural values into AI design. The Partnership on AI's inclusive dialogues, engaging philosophers and social scientists, could reduce biases and align systems with pluralistic ethics (p. 15).

While CPAX compellingly outlines AI's existential risks, I conclude that its premises overstate the likelihood of harm. Intelligence explosions are not inevitable, and instrumental convergence depends on AI's final goals. This does not negate AI's dangers but reframes them as challenges requiring proactive solutions. By investing in technical safeguards, global cooperation, and ethical frameworks, I believe humanity can harness

AI's potential while mitigating risks. Future research should prioritize empirical assessments of AI's trajectory, ensuring theoretical risks inform rather than paralyze progress toward beneficial innovation.

Reference

Vold, Karina, and Daniel R. Harris. "How Does Artificial Intelligence Pose an Existential Risk?" *Oxford Handbook of Digital Ethics*, Ed. C. Veliz, 2021, pp. 1–34.