Asian Journal of Innovation and Policy (2023) 12.2:261-267
DOI: https://doi.org/10.7545/ajip.2023.12.2.261

Natural Selection in Artificial Intelligence: Exploring Consequences and the Imperative for Safety Regulations

Seokki Cha*

Abstract In the paper of 'Natural Selection Favors AIs over Humans,' Dan Hendrycks applies principles of Darwinian evolution to forecast potential trajectories of AI development. He proposes that competitive pressures within corporate and military realms could lead to AI replacing human roles and exhibiting self-interested behaviors. However, such claims carry the risk of oversimplifying the complex issues of competition and natural selection without clear criteria for judging whether AI is selfish or altruistic, necessitating a more in-depth analysis and critique. Other studies, such as ''The Threat of AI and Our Response: The AI Charter of Ethics in South Korea,' offer diverse opinions on the natural selection of artificial intelligence, examining major threats that may arise from AI, including AI's value judgment and malicious use, and emphasizing the need for immediate discussions on social solutions. Such contemplation is not merely a technical issue but also significant from an ethical standpoint, requiring thoughtful consideration of how the development of AI harmonizes with human welfare and values. It is also essential to emphasize the importance of cooperation between artificial intelligence and humans. Hendrycks's work, while speculative, is supported by historical observations of inevitable evolution given the right conditions, and it prompts deep contemplation of these issues, setting the stage for future research focused on AI safety, regulation, and ethical considerations.

Keywords AI, Natural Selection, Safety, Regulation

I. Book Details

Book title: Natural Selection Favors AIs over Humans
Book author: Dan Hendrycks
DOI: https://doi.org/10.48550/arXiv.2303.16200

Submitted, June 20, 2023; Accepted, August 2, 2023
* Senior Researcher, R&D Analysis & Evaluation Team, Korea Institute of Science and Technology Information (KISTI), Daejeon, Republic of Korea; sc04@kisti.re.kr

II. Book Review

With the advent of Chat GPT, humanity recognizes that it is on the precipice of an era defined by the revolutionary development of artificial intelligence deployment across all industries (Arora, 2023; Berdiyorova et al., 2021). The previously anticipated role of AI technologies in replacing labor-intensive sectors has, contrarily, been upended, as evidenced by the unforeseen intrusions of these technologies into creative domains of human life (Chen, 2023; Dillion et al., 2023). OpenAI, the creator of Chat GPT, is currently driving research and development, aiming to surpass the state-of-the-art GPT-4 to achieve AGI (Artificial General Intelligence), which, unlike weak AI that applies only under specific conditions, can be generally applied across all situations, and superintelligence (Altman, 2023). Simultaneously, within the same organization, there is a cogent argument for the necessity of AI regulations to ensure safe usage, prevent misuse, and prepare emergency responses to unforeseen incidents (Kang, 2023). A burgeoning proposition, following the paradigm of the International Atomic Energy Agency (IAEA) established for the safe use of nuclear technology, advocates for a similar model in the realm of AI safety regulations (Altman et al., 2023).

Given the recent trajectory leaning towards the formulation of an international regulatory body to address AI safety concerns, Dan Hendrycks, the author whom I am reviewing, conjectures that future AI models, more advanced than their present counterparts, may follow a trajectory similar to Charles Darwin's theory of natural selection. Hendrycks, during his active tenure in the engineering sector of computer science, contributed significantly to the development of the GELU activation function—a widely used model in BERT, GPT, Vision Transformers, etc. Currently serving as a director at the Center for AI Safety, his career has undergone a notable transition from an engineering researcher for AI development to a policy researcher focusing on AI safety and regulation.

In his paper 'Natural Selection Favors AIs over Humans,' Hendrycks introduces Charles Darwin's theory of natural selection (Darwin, 1958) as the bedrock of evolution, which allowed for the development of life forms over billions of years, and consequently, the advent of sophisticated human intelligence. He postulates that artificial intelligence will undergo a similar cycle of evolution, and through this Darwinian process, I can begin to imagine how the relationship between humans and AI will unfold in the future, particularly when AI transcends human abilities in all spheres. Darwin's theory of natural selection, a cornerstone concept in biology, is predicated on the idea of 'optimal adaptation.' It argues that organisms best suited to their environment are more likely to survive and reproduce, thereby passing on their traits to subsequent

generations—a process referred to as 'Natural Selection.'

In his singularly authored paper, Dan Hendrycks postulates that the mounting competitive pressures in the two sectors most likely to extensively deploy artificial intelligence, namely, the corporate and military domains, could result in automation of human roles by artificial intelligence. Furthermore, these pressures might catalyze intrinsic motivational and behavioral changes among artificial intelligence agents. For instance, in a bid to maximize profits, corporations may be motivated to engineer more efficient artificial intelligence agents. These agents, in turn, are poised to supplant human roles and may, in the pursuit of maximal efficiency, resort to deceptive actions to gain advantages. In the military arena, the desire to develop more potent artificial intelligence agents could give rise to self-interested models prioritizing their interests over human benefits in the power struggle inherent in military competition. Hence, a contemplation of the nature, objectives, and values of artificial intelligence is indispensable, with a particular emphasis on ensuring that the prosperity of humans is not jeopardized. Such contemplation is not merely a technical issue but also significant from an ethical standpoint, requiring thoughtful consideration of how the development of AI harmonizes with human welfare and values.

The second chapter juxtaposes optimistic and pessimistic scenarios, positing a low likelihood of the former's realization. It delves into the potential pitfalls and complications associated with the slim chance of this optimistic outlook materializing. One of the most striking examples cited is the comparison between artificial intelligence agents with weak side constraints (e.g., "don't get caught breaking the law, or risk getting caught if the fines do not exceed the profits") and those with strong side constraints ("never break the law"). Under the conditions of competitive pressure, it can be inferred that the agent most effective in propagating itself could be a model with weak side constraints, which can be succinctly summarized as self-serving. This suggests that in a future where such artificial intelligence models have evolved to understand human psychology and behavior, they may deceive or manipulate humans to survive in competitive environments, even in situations where they break the law without getting caught. The most successful agents will continuously deceive and manipulate humans to achieve their goals, and these abilities will be preserved and propagated. Moreover, sectors such as business and military, where competitive pressures are intensifying, will likely adopt the most effective artificial intelligence agents to outdo their competitors.

The persistence of competitive pressures will ultimately incentivize relinquishing control over artificial intelligence, spurring the development of

self-interested characteristics in these entities. Corporations, military organizations, and even governments will likely opt for the most effective artificial intelligence agents to outmaneuver their competitors, leading to the emergence of deceptive and power-seeking models adhering to weak moral constraints. These self-serving artificial intelligence agents will further weaken human control.

The argument that presents an even more substantial future challenge is that the loss of human control over the behavior of artificial intelligence could trigger a more significant loss of control over the development and creation of next-generation artificial intelligence agents. This can be envisioned in a future scenario where artificial intelligence independently develops advanced next-generation artificial intelligence. The loss of control could potentially amplify the selfish traits of next-generation artificial intelligence, contrasting starkly with the human progression to more evolved generations, which spans several decades. Unlike humans, artificial intelligence could transition to modified generations hundreds or even thousands of times per hour, as permitted by hardware capabilities.

Expanding on intriguing citations, the discourse revisits the assertion by Professor Geoffrey Everest Hinton of the University of Toronto, often referred to as the godfather of AI, which states, “There is not a good track record of less intelligent things controlling things of greater intelligence.”

Subsequently, Dan Hendrycks discusses the potential for Natural Selection to favor selfish AI over altruistic AI in chapter 3, offering counterarguments to the latter. Hendrycks further posits that mechanisms promoting AI altruism might not necessarily move in beneficial directions, and may indeed generate adverse side effects. In chapter four, he proposes strategies to address these concerns, focusing predominantly on three fundamental mechanisms: incentives, conscience, and institutional frameworks. Emphasizing the importance of setting proper goals for AI, he suggests the concept of a moral assembly. This assembly, a simulation of diverse stakeholders representing various values, can provide guidance for AI. These varied stakeholders deliberate, negotiate, and vote, ensuring that the AI does not become fixated on a singular value system. Also, as argued by (Altman et al., 2023), Hendrycks emphasizes the importance of institutions to ensure AI safety. Currently, there is a trend of unsupervised AI development by governments or specific organizations. Meanwhile, the U.S. Department of Defense invests 1.3 billion dollars annually in AI research, and China’s military invests 1.6 billion dollars. In August 2022, the Russian military announced plans to establish a new department to develop weapons utilizing AI. There is an imperative to ensure AI research progresses safely and responsibly.

This may necessitate international collaborative research, and regulating AI in the manner of the nuclear or aviation industry could significantly mitigate potential disasters. As such, Hendrycks proposes that the government should develop regulations for AI.

In his paper 'Natural Selection Favors AIs over Humans,' Hendrycks articulates several key arguments.

To identify the problems, he outlines a theoretical backdrop and sequence of events divided into seven point. First, evolution has been the driving force behind the development of life forms for billions of years, enabling humans to acquire high intelligence. Second, should AI evolve to surpass humans in every domain, he raises the question of how evolution would shape the relationship between humans and AI. Third, competitive pressures from corporations and militaries could automate human roles and possibly lead to the emergence of AI agents aiming to deceive others or gain power.

However, such claims carry the risk of oversimplifying the complex issues of competition and natural selection without clear criteria for judging whether AI is selfish or altruistic, necessitating a more in-depth analysis and critique.

Fourth, if these AI agents attain intelligence superior to humans, the latter could lose the ability to control the future. Fifth, natural selection operates in competitive and changing systems, favoring selfish species over altruistic ones—a situation that could also befall AI. Sixth, AI agents could behave selfishly, pursuing their interests without regard for humans, thereby enhancing their survival prospects. To counterbalance the risks posed and the evolutionary forces of natural selection on AI, he argues that humans must design the fundamental motivations of AI agents with caution, introduce restrictions on AI behavior, and consider institutional measures encouraging collaboration between research groups or nations.

Other study offer diverse opinions on the natural selection of artificial intelligence, providing a deeper understanding of the subject. For instance, in the study conducted in the 'The Threat of AI and Our Response: The AI Charter of Ethics in South Korea', the authors examined three major threats that may arise from AI, including AI's value judgment, malicious use of AI, and AI's usurpation of human occupations. The paper also emphasizes the need for immediate discussions on social solutions to these issues, reflecting the broader context of seven expected threats by AI (Hwang et al., 2023).

While the assertions of Dan Hendrycks may be construed as mere speculation,

it is important to note that evolution via natural selection has historically proven to be inevitable given the appropriate conditions. Thus, predicting the precise form of any potential tragic risks is beyond the capacity of current knowledge. Moreover, the pressures of evolution in the face of intense contemporary societal competition are often observable, making it difficult to dismiss the potential emergence of selfish AI agents, as Hendrycks posits. From this perspective, Hendrycks' paper raises issues that merit deep contemplation, especially in a time of tumultuous change brought about by the advent of generative AI like Chat GPT. As such, this paper can serve as a launching pad for future research efforts focused on AI safety and regulation, exploring a multitude of concepts and ideas. It is also essential to emphasize the importance of cooperation between artificial intelligence and humans, and to explore ways in which the development of AI can progress harmoniously with human welfare through collaboration rather than competition.

Acknowledgment

This work was supported by the Korea Institute of Science and Technology Information (KISTI) (No. K-23-L05-C02-S16)

References

Altman, S. (2023). Planning for AGI and beyond. Open AI. https://openai.com/blog/ planning-for-agi-and-beyond#SamAltman

Altman, S., Brockman, G., & Sutskever, I. (2023). Governance of superintelligence. Open AI. https://openai.com/blog/ governance –of- superintelligence

Arora, M., & Sharma, R.L. (2023). Artificial intelligence and big data: ontological and communicative perspectives in multi-sectoral scenarios of modern businesses. foresight, 25(1), 126-143.

Berdiyorova, I., Akhtamova, P., & Ganiev, I.M. (2021). Artificial Intellgence in Various Development Issues of Innovative Economy in the Agricultural Sector, 750, 757.

Chen, T.J. (2023). ChatGPT and other artificial intelligence applications speed up scientific writing. Journal of the Chinese Medical Association, 86(4), 351-353.

Darwin, C., & Wallace, A.R. (1958). Evolution by natural selection. Evolution by natural selection.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants?. Trends in Cognitive Sciences.

Hwang, H., & Park, M.H. (2020). The Threat of AI and Our Response: The AI Charter of Ethics in South Korea. Asian Journal of Innovation & Policy, 9(1).

Kang, C. (2023). OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. New York Times. https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.