Ayhan Okuyan
Barış Akçin
Mustafa Bay
Berkan Özdamar
Deniz Doğanay

YANN LECUN

In his interview, Yann Lecun states that a Bostrom-style approach to AI is far-fetched, meaning that the AI will always be controllable in the course of its evolution. Bostrom discusses the possibility of an AI system that can eliminate the boundaries of human ethics and harm humans to reach its goal. Lecun, on the other hand responses with the argument that the learning algorithms used to declare the goal of performance of AI systems are run by minimizing mathematical objective functions, and proposes that for an AI, to follow its own agenda somehow, an objective function should be defined with intrinsic motivations. Then he continues with the proposition that this structural form is opposite of how the objective functions are present in humans, hence, cannot be applied to machines since we are imitating organic neural systems to develop the artificial AI models.

We disagree with the Lecun's opinions presented for three separate reasons. First, the concept of deep learning is presented to observe cognition on a more abstract level, such that the algorithm learns by itself without full supervision on its weights. In a neurological perspective, we are unable to understand what the deeper levels of the network, segments under the cortex, hold as information. Hence, in our designed replicas, we also mostly do not understand what features are extracted. Hence, in a designed objective function, there is a possibility for the algorithm to diverge to a path that was not foreseen due to unexpected data or the design itself. As a result, we can say that there is a possibility that although the technology to contain positive values, the outcome can be harmful.

Secondly, we believe that our intrinsic objective functions are not fully learned as Lecun suggests. Most objective functions in human brain work on a punishment-reward emphasis, meaning that the action is taken based on the predicted reward. However, there are some observed biases that overpower these learned decision-making algorithms in critical times when the individual is under pressure. The human brain is a highly complex structure that we are not currently in the position to declare causal relationships and make hard conclusions on how it operates. Furthermore, Lecun states in the interview that, they "don't know if the brain minimizes an objective function" [4], presenting they don't have enough knowledge on if and how brain prioritizes learning one feature over the other.

On the other side, there are countless AI experts and scientists that believe in the Bostrom-style thinking. One most famous example of such scientist is the CEO of OpenAI, Elon Musk. Musk states in an interview:

> ***"The rate of improvement is really dramatic. We have to figure out some way to ensure that the advent of digital super intelligence is one which is symbiotic with humanity. I think that is the single biggest existential crisis that we face and the most pressing one."*** [1]

He claims that the probability of a super intelligence, meaning an intelligent model that has high cognitive abilities and general, can create the biggest existential crisis the humanity has faced. Unlike what Lecun describes, Musk believes that AI has the capacity to remove human out of the equation by becoming a higher form of intelligence. Some evolutional psychologists believe that the only reason human has dominated other species is due to its distinctive capabilities related with their brain

operations [2] and if the super-intelligence turns to reality, then there exists a possibility that event would trigger human-extinction. As an example, we can propose AlphaGO, which is an algorithm that plays Go, the most mathematically complex board game ever created. AlphaGO algorithm has trained for only three days before reaching a super-human level [3]. This result shows that there is possibility for AI to reach these levels of intelligence realistically.

Overall, we find Lecun's explanation of AI not becoming a viable threat for human life enough for three reasons. First one is that this field, at this moment, is not comprehensive enough to understand what the learning algorithm perceives in the deeper layers; hence we cannot make deterministic comments about what the algorithm precisely learns. Then, we disagreed with Lecun as he facilitated the idea that the intrinsic learning functions in humans are all learned-value functions. We believe there are many cases that the decisions are overwhelmed by more hard-wired biases, like reflexes and we cannot say for sure that this is impossible during the training of AGIs (Artificial General Intelligences). Furthermore, there are various scientist that accredit Bostrom's ideas, claiming that super intelligence does indeed have the capacity to drive humanity into extinction.

The segment of the interview that we have disagreed on is given below [4].

YANN LECUN: Well, there is the issue of objective function design. All of those scenarios assume that somehow, you're going to design the objective function—the intrinsic motivations—of those machines in advance, and that if you get it wrong, they're going to do crazy things. That's not the way humans are built. Our intrinsic objective functions are not hardwired. A piece of it is hardwired in a sense that we have the instinct to eat, breathe, and reproduce, but a lot of our behavior and value system is learned.

Ayhan Okuyan
Barış Akçin
Mustafa Bay
Berkan Özdamar
Deniz Doğanay

References

[1] CatClifford, “Elon Musk: 'Mark my words - A.I. is far more dangerous than nukes',” *CNBC*, 14-Mar-2018. [Online]. Available: https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html. [Accessed: 24-Dec-2019].

[2] C. to W. projects, “Hypothesis that Artificial General Intelligence could result in human extinction,” *Wikipedia*, 02-Dec-2019. [Online]. Available: https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvRXhpc3RlbnRpYWxfcmlza19mcm9tX2FydGlmaWNpYWxfZ2VuZXJhbF9pbnRlbGxpZ2VuY2U. [Accessed: 25-Dec-2019].

[3] J. Whittlestone, A. Weller, S. Hawking, and B. Barnes, “Risks from Artificial Intelligence,” *University of Cambridge Centre for the Study of Existential Risk*. [Online]. Available: https://www.cser.ac.uk/research/risks-from-artificial-intelligence/. [Accessed: 24-Dec-2019].

[4] M. Ford, *Architects of intelligence: the truth about AI from the people building it*. Birmingham, UK: Packt Publishing Ltd, 2018.