34TH BLED ECONFERENCE
Digital support from crisis to progressive change

University of Maribor Press

Make Us Smile! AI and the Violation of Human Intentions

CHRISTOF WOLF-BRENNER

Abstract In his book Superintelligence, Nick Bostrom points to several ways the development of Artificial Intelligence (AI) might fail, turn out to be malignant or even induce an existential catastrophe. He describes 'Perverse Instantiations' (PI) as cases, in which AI figures out how to satisfy some goal through unintended ways. For instance, AI could attempt to paralyze human facial muscles into constant smiles to achieve the goal of making humans smile. According to Bostrom, cases like this ought to be avoided since they include a violation of human designer's intentions. However, AI finding solutions that its designers have not yet thought of and therefore could also not have intended is arguably one of the main reasons why we are so eager to use it on a variety of problems. In this paper, I aim to show that the concept of PI is quite vague, mostly due to ambiguities surrounding the term 'intention'. Ultimately, this text aims to serve as a starting point for a further discussion of the research topic, the development of a research agenda and future improvement of the terminology.

Keywords: • Artificial Intelligence • Digital Ethics • Reinforcement Learning • Control Problem • Perverse Instantiations •

CORRESPONDENCE ADDRESS: Christof Wolf-Brenner, e-mail: christof.brenner@gmx.at

1 Introduction

Recently, the importance of being in control of what Artificial Intelligence (AI) does has moved to the center of attention. There appears to be a broad consensus that AI should not do what we homo sapiens do not want it to do. The concept of Perverse Instantiations (PI) consequently describes cases where AI succeeds to achieve goals but does so in violation with our intentions. For instance, best-selling author and philosopher Nick Bostrom brought forward the thought experiment of AI choosing to paralyze human facial muscles into constant smiles to achieve the goal of making humans smile. Of course, similar more or less realistic cases can be constructed for various domains. AI might for instance attempt to achieve the goal of reducing maternal mortality by sterilizing all male homo sapiens, or to improve schoolchildren's grades by providing them with the answers to the next test beforehand. While ultimately, the goal is achieved in each scenario, the way it was achieved was not intended.
According to Bostrom, PI ought to be avoided because of the violation of human designer's intentions. In this paper, I aim to show that the current concept of PI is quite vague, mostly due to inaccuracies surrounding the term 'intention'. The prevailing terminology, if we took it seriously, would force us to label most ways of achieving goals that were uncovered by AI as unintended and consequently, as PI. Ultimately, this text serves as a starting point for a further discussion of the research topic and aims to provide reasoning why we should look deeper into the matter at hand.

2 The Meaning of 'Intention'

At first glance, Bostrom's definition of PI as AI *"discovering some way of satisfying the criteria of its final goal that violates the intentions of the programmers who defined the goal"*[1] sounds reasonable. In essence, he claims that AI going against the intentions of its designers might lead to undesirable outcomes. But what is the real meaning of 'intention' that he has in mind when framing his notion of PI? Revisiting the thought experiment of producing smiles, Bostrom explicates that violating intentions is to be understood as *"not to do what the programmers meant when they wrote the code that represents*

[1] Bostrom 2017, p. 146

this goal"[2]. But what *did* they mean when they tasked AI with coming up with ways to make humans smile? To answer that, we will take a short detour into the realm of Reinforcement Learning (RL) as a state-of-the-art example and the technologically simplest instance of AI that Bostrom himself uses when describing PI.

RL is a subfield of AI and Machine Learning that is focusing on automatically learning optimal decisions over time. For simplicity's sake, imagine a mouse-in-a-labyrinth kind of experiment. Within the maze, the designer randomly places non-deadly traps and delicious food. A mouse, conceptually referred to as an RL agent[3], is placed in the maze and can perceive its environment through its senses. Of course, its objective is to try and obtain as much food as possible without getting hurt by the traps.[4]

To achieve its goal, the digital rodent can mix and match actions from a finite list of actions called the *action space*, i.e., turn around, move, wait, gnaw, jump etc. RL can then be understood as repeatedly putting the same mouse into very many mazes to finally make it learn to automatically choose the optimal combination of actions from its action space to maximize the aggregate reward, i.e., eat the maximum amount of food while stepping into the fewest traps.[5]

3 Perverse Instantiations Emerging from Underspecified Goals

If we were to transfer the concept of RL to Bostrom's example of tasking AI to make us smile, we encounter a few challenges. Initially, we would need to adequately represent the goal in a way so that we can give our agent feedback on how well it has done, which includes translating 'make us smile' into a form that the AI can understand. First, it is up for interpretation what counts as a smile. Second, it is unclear whether the AI is meant to achieve the maximum number, duration, intensity etc. of smiles. Third, the term 'us' is ambiguous and contextual, potentially leaving us with an AI that might produce a lot of smiling corpses.

Making humans smile by paralyzing facial muscles might not be intended, but if what is intended is not made explicit in the goals, a RL agent will not be able to take it into account, as it is only interested in maximizing goal achievement, thus reward gain.

---

[2] Bostrom 2017, p. 147
[3] I use the term 'agent' here in the sense in which it is used in the domain of Reinforcement Learning.
[4] See Lapan 2020, pp. 1–5
[5] Ibid.

As shown with Bostrom's own example, it is in fact not the case that AI discovered a loophole to perversely instantiate the sought-after goals. The problem appears to be the programmers' vague formulation of goals.

Yet, if there is so much room for interpretation, how can we then speak of a violation of intentions? If I kindly ask you to prepare a sandwich for me, and you go ahead and fix up a ham and cheese sandwich, can we really speak of you violating my intentions because unbeknownst to you, I am a vegetarian? It is my firm believe that we should not put the blame on the sandwich-maker if he only knew half of what was truly expected of him.

4 Perverse Instantiations Emerging from the Way Goals Are Achieved

Let us, however, assume, that there are cases in which the goals are in fact stated in a way such that the intentions of the programmers are absolutely and unmistakably clear. Could there still be PI? Bostrom's examples strongly imply that even clear-cut goals can be perversely instantiated by employing unintended ways to reach them.

Reconsidering our mouse-in-a-labyrinth scenario, the finite list of actions that the mouse can take to interact with the maze is called the action space. It can be understood as a kind of toolbox that programmers infuse their RL agents with to act in the environment.

However, there are two ways I can think of by which a finite list of actions might still circumvent the designer's intentions. First, knowing all individual actions might simply not be sufficient to check for conformance. For instance, if our mouse in the maze would be able to act in three distinct ways, a solution to optimize reward gain might include chaining together these actions up to ten times, which already amounts to well over 500 possible combinations. There is a significant chance for combinations that the programmers never would have thought of, and therefore, could not have been intended by them.

Second, unexpected interactions between agents and complex environments in RL leave ample room for PI. Researchers tasked RL agents to play hide and seek against each other, and the emerging collaborative strategies far exceeded what was initially anticipated. To win, seekers learned to surf on crates by exploiting the way movement was implemented. By doing so, they were able to overcome the shelters that hiders had built as part of their defensive strategy[6]. The sheer range of things

[6] See Baker et al. 2020, p. 6

that agents could do in the game world was simply ungraspable for humans beforehand, which lead to unintended solutions, even though the goal was never unclear or misunderstood.

5 The Intention of Violating Intentions

By design, RL scenarios aim to produce optimal solutions for gaining some reward with minimal or no explicit instructions on how to do so. In a way, RL agents having the freedom to experiment within some boundaries is exactly what we intend to do when employing that kind of AI. So, following our arguments from the previous sections, should we label every instance of AI not doing what the programmer meant as PI?

When Lee Sedol, one of the best Go players on the planet, sat down to play against AlphaGo, an AI trained by way of RL, the human champion lost four out of five rounds. Move 37, which the AI came up with in the second game, stunned Sedol. In thousands of years of humans playing Go, nobody had ever come up with something as inhuman, unique, or creative.[7] But what were the intentions of the designers of AlphaGo? Clearly, they meant to design an AI that can play and excel at Go. Obviously, Move 37 was intended insofar as it is in accordance with the goal, which I loosely interpret as win at Go by playing the game by the rules. But did the programmers intend Move 37? Arguebly, they did not. Move 37 was unexpected for the opponent, the spectators and even more so for the creators of the AI. AlphaGo itself estimated that a human player would have played this move with a probability of one in 10,000 but decided to go for it anyway [8].

Bostrom does slightly hint at his account of intention being tied to the way goals are achieved in the sense of a method, and not the goals themselves[9]. If this is the case, labeling cases as PI boils down to the question whether we can at the same time intend to search for the optimal way of satisfying a goal and already know the result of the search. Searching for a solution and already intending it seems contradictory. However, if we tasked AI to make us smile, play Go or eat food and avoid traps, any case of it achieving these goals through unintended ways would still have to labeled as PI. It appears that simply reducing PI to unintended ways to achieve goals is not enough to clearly explain the problem arising from applying AI in such scenarios.

[7] See Holcomb et al. 2018, p. 68
[8] See Holcomb et al. 2018, p. 70
[9] See Bostrom 2017, p. 147

6 Conclusion

In this paper, I have shown that the concept of PI hinges on the definition of 'intention' and consequently, what the intention latches on to. I have concluded that PI arising from underspecified goals are simply a matter of intentions not being laid out to AI in an understandable, unmistakable, and complete manner. By no means would we be justified to declare our creation at fault, because everything it does is directed at what it knows about our intentions, expressed through goals, and achieving them.

Additionally, I have discussed that even for cases in which intentions are perfectly clear to AI, there still is room for PI based on actions and interactions with the environment not being fully understood by designers beforehand. It appears contradictory to claim to have intended a particular way of achieving a goal which AI just now discovered. In contrast, not every time some goal is achieved in a way that designers were previously unaware of is unintended, and thus a case of PI.

It is clear that the definition of 'intention' plays a pivotal role in identifying PI. However, as I have shown, there is more work to be done on the terminology. First, a starting point for further conceptual contribution to the topic could be an investigation of a potential conflation of the terms 'unintended' and 'unanticipated' in the context of unintended consequences. This distinction could shed some light on the issue of mislabeling cases as PI.

Second, empirical research could be conducted with regards to what cases of AI, humans or other animals achieving goals through unintended ways people would in fact label as PI. Consequently, bias for or against our artificial creations and fellow planet dwellers could be identified and elaborated on.

References

Baker, B.; Kanitscheider, I.; Markov, T.; Wu, Y.; Powell, G.; McGrew, B.; Mordatch, I. (2020): Emergent Tool Use From Multi-Agent Autocurricula. In: 8th International Conference on Learning Representations. Addis Ababa, April 26-30, 2020.

Bostrom, N. (2017): Superintelligence. Paths, Dangers, Strategies. 2nd ed. Oxford: Oxford University Press.

Holcomb, S. D.; Porter, W. K.; Ault, S. V.; Mao, G.; Wang, J. (2018): Overview on Deepmind and its AlphaGo Zero AI. In Proceedings of the 2018 International Conference on Big Data and Education, pp. 67–71.

Lapan„ M. (2020): Deep Reinforcement Learning Hands-On. 2nd ed. Birmingham: Packt Publishing.