There's (no) harm in trying.

In March 2016, Microsoft's chatbot 'Tay' developed racist and homophobic behavior in less than a day after its release.[1] Only two months later, an AI system used by the US court to assess the risk of prisoners committing a future crime wrongly marked black offenders significantly more often than those of other races.[2] Similarly, the robot evaluating the photographs of participants in "The First International Beauty Contest Judged by Artificial Intelligence" exclusively crowned white winners.[3] In an on-going stream of AI incidents, these examples indicate that AI systems struggle to act ethically but are, in fact, prone to bias. This raises a crucial question: Are we able to prevent intelligent agents' susceptibility to bias?

For decades, AI has been developed focusing almost exclusively on technological and economical benefit. The term 'intelligence' itself states, however, that these systems demonstrate an intellectual capacity that goes far beyond data processing – they are programmed with algorithms and continuously fed with data, learn from experiences they make and eventually adjust their performance based on those. The idea sounds promising but has a weak point: A machine does not have an initial sense of morality but is taught how to act. Yet, at the moment it is free to adjust its behavior, its evolution becomes opaque. And so far, we have done little to gain control over this.

Any AI system that is programmed to implement tasks that require intellectual capacities should necessarily be taught to do so in an ethical and unbiased way. Consequently, some sort of ethical principle has to be fed into the algorithm of a machine. If these can be specifically stated, acting ethically is merely a matter of following guidelines – the system would simply compute whether its actions are allowed by the rules.[4] Every student who took at least a Philosophy 101 class should now frown and point out at least three problems:

First, which code of ethics are we talking about? Utilitarianism? Kant's categorical imperatives? Aristotle's virtue ethics? The bible? Second, if an AI system is guided by principles, can it still be seen as unbiased? And third, how can anyone ensure that the rules are followed at any time even though the machines are allowed for autonomous behavior?

In order to clarify these questions, I want to introduce two different approaches to artificial morality offered by Wendell Wallach and Colin Allen: The top-down approach, which is rather strict and merely takes a set of rules that are turned into an algorithm and fed to the machine. For example, a computer could compute the consequences of its actions in a utilitarian way in order to rank them morally. Or it could be taught to only implement actions that are based on a morally right motive (which obviously has to be stated, too) in disregard to its outcome. Clearly, such an ethical code would have to be chosen wisely and made a universal law in the developing process of AI. However, while the top-down approach seems reasonable as it provides solid, legal guidelines and would simplify the

question about accountability, too, it fails to exhaust the potential of AI system as learning machines. The bottom-up approach, on the other hand, is more open as it promotes exploration, learning and reward for praiseworthy behavior.[5] It allows for individual judgments that take previous experiences, predictable outcomes and the framework conditions into account. At the moment, AI systems that are designed as learning machines follow at least partly the bottom-up approach but lack a base to develop their conduct on. A combination of both approaches would be desirable as it suggests more control over the direction of development but an openness towards individual adjustments depending on the situation.

Nevertheless, this is a source of potential risks for an undesirable evolution of the AI system. The very fact of following predetermined rules excludes complete impartiality.[6] I must confess that this seems petty, as the existence of such a thing as complete impartiality is debatable anyways, but who can ensure that no bias has been put into the algorithm either accidentally or purposefully? In practice, we simply don't know what the AI system is going to do with the data eventually – the best intentions could inadvertently turn into biased results.[7]

Bias, as defined in the Oxford English Dictionary, is an "inclination or prejudice for or against one person or group, especially in a way considered to be unfair".[8] However, the borders are blurry here since one can question if this is actually bias if a machine evaluates data and draws conclusions regarding probable outcomes. In this sensitive area, it is possible that we are biased ourselves, hoping for the AI machines to give us the most neutral, average and, in times of doubt and anxiety, reassuring result. Perhaps because we ourselves fail to assess anything without embedding it in a historical, cultural, and social context. For this reason, we need AI systems that we can fully rely on. That we can be absolutely certain about them acting neither to satisfy any expectations, nor to confirm any prejudices or, in the worst case, to their own advantage.

While for the longest time AI research was barely restricted, a bill, called the "FUTURE of Artificial Intelligence Act of 2017", just recently has been referred to the US Committee of Commerce in order to place legal guidelines around the usability and evolution of artificial intelligence. On a total of 14 pages, the word *ethics* is mentioned only twice: The demanded Committee shall provide advice regarding Ethics training and development for those working on AI. Additionally, the members of the Committee should show expertise in, inter alia, ethics.[9] That's it. This long overdue act, as much legal guidance as it suggests apart from this, should make anybody with actual human intelligence feel uneasy as it primarily shows how little we know about the actual consequences of our research. Neither is a code of ethics defined nor does it make any suggestions as in how to include such in the developing process. The proposed diversity in workforce is definitely necessary, but far from sufficient. Before we proceed in giving free hand to our machines, we should clarify all the questions mentioned so far. AI has already found its way in our

courts, cars, and homes, and we have an almost blind faith in its functioning. Who knows though, if the break assistant eventually decides not to slow down for a jaywalker because it recognizes him or her as a former criminal? Or worse, because of his or her race? Engineers, physicians, mathematicians and many more – they all have been working hard to get to the point we are right now. Incredible progress in technology has been made, objects whose functions we could not even imagine a decade ago have been successfully developed – by careful conceptualization and playtesting. Realizing new dreams necessarily means leaving the comfort zone, taking a risk and trying something that no one has ever tried before. But if no certain predictions can be made about the outcome, such a risk could also be regarded as madness. In this case, trying could definitely do harm. The vast number of incidents as well as the vague formulated act demonstrate that we are unable to prevent AI systems from developing in an undesirable direction and thus, cannot prevent susceptibility to bias at this point. My history teacher in high school once exposed me when he read my extensive answers to his exam questions to the class. While they were all correct, they just had little to do with the questions posed. I simply wrote down what I knew because I had no clue about the actual topic. He said, sometimes, it is better to pause until you know how to proceed. I wish somebody would tell this to our AI developers.

End Notes

[1] James Vincent, „ Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day," *The Verge*, March 2016, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist (accessed January 29, 2018).

[2] Stephen Buranyi, „ Rise of the racist robots," *The Guardian*, August 2017, https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses (accessed January 25, 2018).

[3] Sam Levin, „A beauty contest was judged by AI and the robots didn't like dark skin," *The Guardian*, September 2016, https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people (accessed January 29, 2018).

[4] Wendell Wallach, *Moral Machines,* (New York, Oxford University Press, 2009), 83.

[5] Wallach, *Moral Machines,* 85.

[6] Wallach, *Moral Machines,* 110.

[7] Larry Dignan, „ Can AI really be ethical and unbiased?," *ZDNet*, October 2016, http://www.zdnet.com/article/can-ai-really-be-ethical-and-unbiased/ (accessed January 25, 2018).

[8] "bias, adj., n., and adv.". Oxford Dictionnaries. January 2018. Oxford University Press. https://en.oxforddictionaries.com/definition/biasfalse (accessed January 25, 2018).

[9] FUTURE of Artificial Intelligence Act of 2017, H. R. 4625, 115th Cong. (2017), 6, 11.