Artificial stupidity: How can we guard against mistakes?

AI Research and Programming

PRG2006

Xhefri Bala

7843

Modernization, referring to the transitional process through which a society undergoes systematic and progressive change, has had, and continues to have a profound impact on the way humans work and live (Bernstein 1971). Consequently, huge technological advancement has occurred within the last few decades (Marsh 2014). Perhaps one of the most noteworthy changes to the technological realm is the rise of artificial intelligence. Artificial intelligence (AI) is a broad concept by which computers or machines are programmed and trained to perform human-like tasks (Raso, Hilligoss & Krishnamurthy et al. 2018). Many of these human-like tasks typically require human emotion and decision-making skills, and therefore, this tends to spark a debate around the ethical considerations of allowing computers to carry out this type of work. One of the most crucial things to consider with AI, is the way in which we can avoid mistakes and ensure a machine is operating as planned. Difficulties may arise with AI in terms of their lack of human emotion and empathy. Moreover, although AI may be very intelligent, their programming may allow them to confuse stimuli and consequently operate in a counter-productive way. Additionally, if AI does malfunction, if can be very detrimental in terms of safety, but also in terms of liability and other legal factors. As a result, it is integral that humans keep learning and developing ways to minimise and avoid these mistakes in order to live harmoniously with AI.

There are a variety of risk factors associated with creating systems that emulate a human behaviour. The difficult part, though, is derived from the fact that humans and AI machines have one overwhelming difference; humans are able to feel a wide range of emotions and robots, arguably, are not (Brookhouse 2020). When making an important decision, humans use those emotions to regulate their behaviour and adapt to situations accordingly. Even the most sophisticated AI systems simply do not have this ability, at least not to the extent that humans do (Brookhouse 2020). Consequently, when AI is programmed to carry out a task, it will carry that task out without eliciting an emotional response and will do so regardless of how destructive that task may be. This lack of emotional reasoning can lead to mistakes. Self-driving cars are a good example here. If a car is asked to reach a destination as fast as possible, it may fulfil the literal request, but do so in a manner that is unsafe (Tegmark n.d). In this instance, the AI is being obedient and fulfilling a request, but may not be able to use emotional reasoning to fill in the blanks of the programmer's request. As such, it is crucial that a harmonious alignment exists between the programmer's intentions and what they actually program (Tegmark n.d).

When AI malfunctions, it can not only cause frustration, but in some cases, can be extremely dangerous or fatal. In these cases, it is crucial to understand the ethics surrounding liability. If a machine is as intelligent as humans argue they can be, then theoretically, they should always perform at the optimal level. If, for whatever reason, a malfunction occurs, then according to human expectation, this would make the machine liable. On the contrary, if humans are the ones programming and training these machines to operate, then it could also be argued that humans are the ones liable for malfunctions. This becomes a particularly important debate when considering large machinery. For instance, it has been said that most cars will be self-driving within the next few decades (Gessner 2019). This means that cars will be automated and run by AI systems. Research has shown that AI-operated cars can easily be deceived by changes to road signage (Eykholt et al. 2017). For instance, it was shown that by placing stickers on a stop sign, AI can be fooled into misclassifying that stop sign and making a detrimental mistake (Eykholt et al. 2017). If one of these cars is involved in a serious accident, then from an ethical standpoint, the liability could be placed on the car manufacturer, the AI system developer, the AI computer itself, or even the human who occupies the car. A significant amount of work is required to ensure AI is safe, reliable and user-friendly enough for the general population to use (Gessner 2019). If a vehicle is marketed as ‘automated’ or ‘self-driving’, it is incredibly important to ensure that potential mistakes are eliminated so that they carry out what they claim to carry out, and these debates can be avoided altogether. This confusion surrounding liability is complex reason why it is crucial to guard against mistakes when creating AI systems.

When programming machines to complete tasks that humans normally complete, naturally, some issues are bound to arise. As such, more research into the ethical considerations and safety implications is needed before AI becomes even more widely used. AI’s lack of emotional regulation is one of the key areas that should be researched further to avoid catastrophes. Additionally, in an ideal world, programmers and AI researchers would need to train AI systems in a way that is directly in-line with the programmer’s intentions. Both of these areas require development before automated systems are used in everyday life by the general population. By working in these two areas, AI mistakes will be far less frequent, which will consequently reduce the need for legal debates surrounding liability.

Mistakes are inevitable when progressive change is occurring, however, when there is so much riding on AI, mistakes simply cannot happen.

References

Bernstein, H 1971, Modernization theory and the sociological study of development, *The Journal of Development Studies,* vol. 7, pp. 141-160.

Bossman, J 2016, Top 9 ethical issues in artificial intelligence, *World Economic Forum*, viewed 22 February 2021, <https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/>.

Brookhouse O 2020, Can artificial intelligence understand emotions? *Telefonica,* viewed 26 February 2021, <https://business.blogthinkbig.com/can-artificial-intelligence-understand-emotions/#:~:text=AI%20and%20neuroscience%20researchers%20agree,and%20emit%20more%20realistic%20emotion>.

Evtimov, I, Eykholt, K, Fernandes, E, Kohno, T, Li, B, Prakash, A, Rahmati, A & Song, D, 2017. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv*, vol. 2, no. 3, p.4.

Gessner, D, 2019, Experts say we’re decades away from fully autonomous cars. Here’s why, *Business Insider Australia,* viewed 27 February 2021, <https://www.businessinsider.com.au/self-driving-cars-fully-autonomous-vehicles-future-prediction-timeline-2019-8?r=US&IR=T>.

Marsh, R, 2014, Modernization theory, then and now, *Comparative Sociology,* vol. 13, no. 3, pp. 261-286.

Raso, F, Hilligoss, H, Krishnamurthy, V, Bavitz, C & Kim, L, 2018, Artificial intelligence & human rights: Opportunities & risks, *Berkman Klein Center Research,* no. 2018-6.

Siva C 2018, Machine learning and pattern recognition, *DZone*, viewed 25 February 2021, < https://dzone.com/articles/machine-learning-and-pattern-recognition >.

Tegmark n.d, Benefits & risks of artificial intelligence, *Future of Life Institute*, viewed 25 February 2021, <https://futureoflife.org/background/benefits-risks-of-artificial intelligence/>.