8-9-2021

On Patching the Moral Vulnerabilities of Artificial Intelligence

Abdelnasser Abdel-Aal
King Faisal University, dr.abdelaal@gmail.com

On Patching the Moral Vulnerabilities of Artificial Intelligence

TREO Talk Paper

Abdelnasser Abdelaal
King Faisal University, Saudi Arabia
aabdelaal@kfu.edu.sa

Abstract

The ethical scandals of artificial intelligence have captured the attention of scholars, experts, regulatory bodies, and policymakers. For example, Rekognition- the Amazon's face recognition system has shown bias against people of color. Similarly, the IBM's Watson oncology software has failed in cancer treatment because designers used poor training data, relied only on the U.S. medical protocols and failed to consider local contexts. Such ethical issues show that Artificial Agents (AAs) have certain ethical vulnerabilities and threats that require mitigation measures. The term ethical vulnerability refers to a fault, flaw, defect, or a shortcoming in the system design that may lead to a misconduct, harm, or bias. Scholars agree that the main reason for these unethical behaviors is the poor data used to train these algorithms.

A key source of moral vulnerabilities is the *agency* of some AAs. Agency means that an AA has the capability to assess situations, set goals, and choose effective strategies to achieve these goals. Currently, organizations use autonomous agents to decide eligibility for loans, target specific individuals for security, recommend an applicant for a specific job, choose taxpayers for audit, and grant visas to foreigners. Such agents should be bounded by certain ethical obligations or human oversight.

Autonomous systems are also ethically vulnerable. These include chatbots, sexbots, and driverless cars. They are also employed for controlling dangerous systems such as atomic reactors, lethal autonomous weapon systems and stock markets. A robot called 'Marty' harassed a woman in the Stop and Shop supermarket in New York. Similarly, a chatbot called 'SimSimi' has been banned in Ireland in 2019 after sexually harassing minors. Therefore, scholars hold that there is a need for human oversight mechanisms to minimize any potential misconducts, bad decisions, or harmful outcomes.

The *intrusiveness* nature of many agents is another issue of concern. Some algorithms (e.g. these of Google, Facebook, recommendation systems) are intuitively intrusive as they are designed mainly to collect, analyze, and utilize private data for recommendations. In so doing, they may not take the consent of data owners. Similarly, drones could be manipulated remotely to infringe on private properties. In addition, persuasive algorithms disseminate misinformation to influence decisions of people with respect to content, political views, or products. Unlike human persuaders (e.g. parents, leaders, teachers, and sales clerks), artificial persuaders usually work covertly and may not honor the moral system of users.

The *non-neutrality* of many AAs influences our actions and decisions. In fact, AI may be the most non-neutral technical force, compared to Internet of things, renewable energy, and biotechnology. Stakeholders are challenged to device instruments to neutralize AAs. Employing *poor or biased data* to train AAs is another ethical flaw or vulnerability. It the main reason for the bias of some algorithms against underrepresented groups. These deficiencies signal calls for designers to consider the moral needs of vulnerable communities and underrepresented groups. One more source of moral hazards is the fact that AAs lack *conscience and sentiment*. Without sentimentality, AAs cannot recognize self-awareness, pain, or pleasure. Absence of these features could be a main source of ethical scandals. The *learning capabilities* of machine learning may also pose ethical hazards. For example, Microsoft had to retire its 'Tay ' bot after disseminating racially prejudiced information about Jews taught by users.

Apparently, new types of threats have emerged which could be called "ethical attacks." These attacks could be conducted innocently by innocent users or maliciously by perpetrators. When a perpetrator discovers an ethical flaw or a vulnerability in the system, he/she/it might exploit it. Thus, there is a need for frameworks for detecting, testing, and fixing these ethical vulnerabilities. The solution could be a suitable ethics patch, a mechanism for human oversight, developing moral algorithms or organizing ethics training for machines to avoid such misconducts or harms.