DOI: 10.24425/academiaPAS.2023.146571

STOKKETE/SHUTTERSTOCK.COM

Piotr Kaczmarek-Kurczak, PhD

is a lecturer at the Department of Entrepreneurship and Business Ethics, Kozminski University, and at Kozminski University's Centre for Space Studies – Kozminski ESA Lab. A member of the Program Board of the Space Entrepreneurship Institute. An outside expert at the Future Industry Platform Foundation in the area of digital business models and applications of technologies of the future. An expert and auditor of ADMA (advanced manufacturing) transformation in industrial companies in Poland.

pkurczak@kozminski.edu.pl

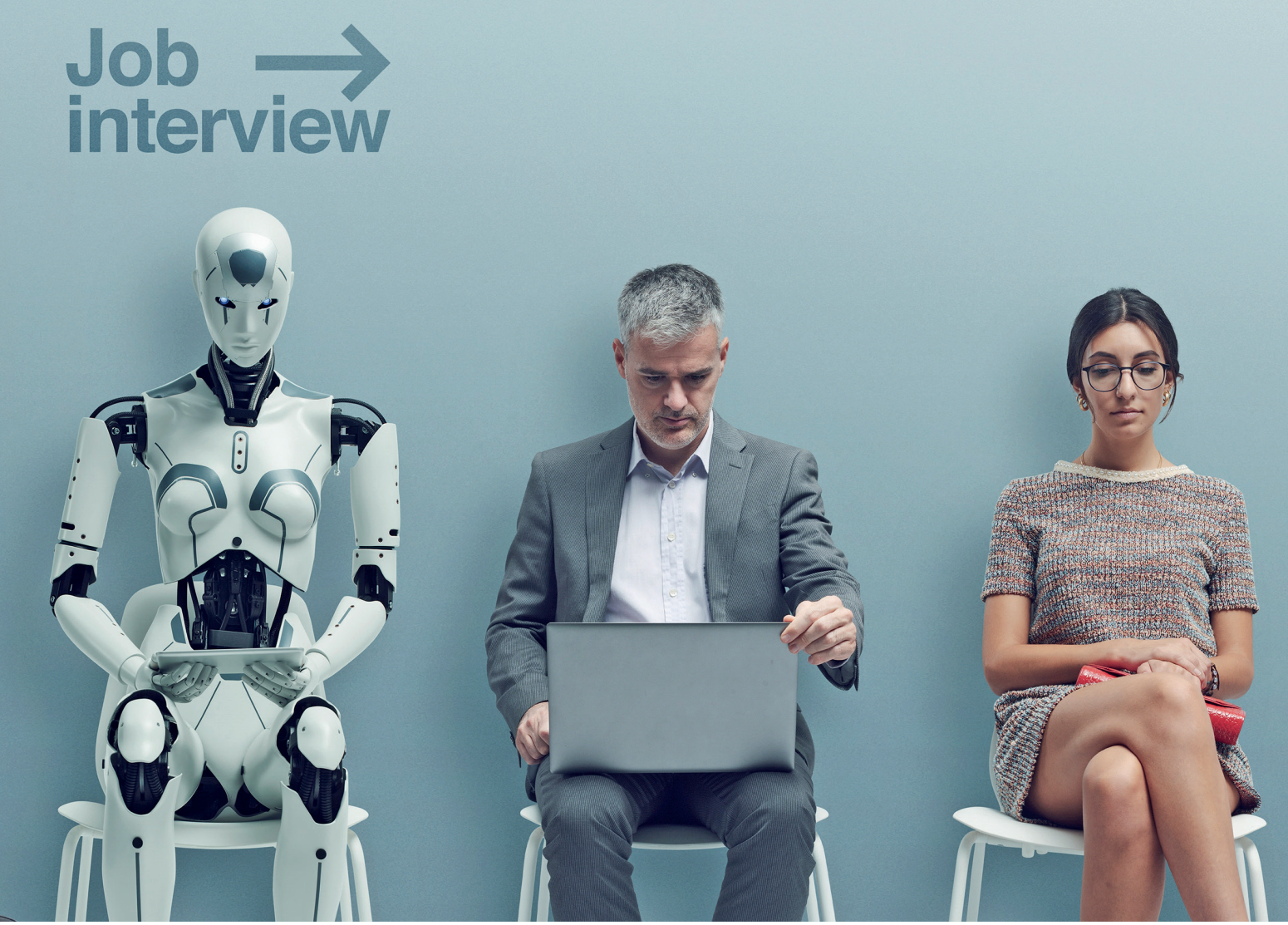
Keeping Artificial Intelligence Under Rein

Rapidly developing artificial intelligence technologies are expected to help us in various sectors of life, but their applications also entail certain risks.

Piotr Kaczmarek-Kurczak

Centre for Space Studies, Kozminski University – Kozminski ESA Lab in Warsaw

Artificial intelligence (AI) is transforming our world in revolutionary ways. The possibilities it open up appear limitless: from autonomous drones to personalized medical treatments. But growing reliance on AI also raises the need to address the emerging safety concerns.



In January 2017, the Shadow RQ-7Bv2, an unmanned aerial vehicle (UAV) used by the US Army for reconnaissance, took off from its launch pad at a training range in southern Arizona in a typical military exercise. Drones of this type, which are typically used for simple terrain observation and guidance of artillery fire, have a range of up to 77 miles, or roughly 120 kilometers, from the ground control station. That particular drone, however, was different: it was fitted with enhanced capabilities to act independently (with autonomy). Once it left the launch pad, the control station lost all communication with the device, which soon began a surprising journey. Immediately after the launch, it changed its course towards the Rocky Mountains and disappeared.

The operators were convinced that the drone would crash against the mountain slopes: it had a limited range and was not designed to operate at high altitudes. But the aircraft somehow managed to rise to an altitude of 4,000 meters and fly across the entire mountain range. It crashed 600 kilometers away from the launch pad, most likely because it had run out of fuel. It remains unclear why the drone performed the whole flight or how it accomplished it. It is suspected that air currents may have played a key role in that they not only extended the drone's range, but also helped it attain the necessary altitude. The direction it chose to take was yet another mystery. Since the reconnaissance unit had been earlier stationed at a base in this part of the country, the device's memory may have retained the coordinates of its former base, and poor Shady – as the press dubbed the drone – was simply "trying to go back home."

Unpredictability

The incident offers a compelling illustration of several key problems arising out of the development of AI and its applications in today's world.

The first of these is uncertainty. In simple machines, like a bicycle or a swing, it is easy both to understand their mechanism and to predict their behavior to some extent (in everyday life, nothing is entirely deterministic, so the element of chaos can never be ruled out completely). However, even machines based on such relatively simple operating procedures as algorithms very quickly generate higher levels of uncertainty – each additional rule increases the complexity of the whole system and the risk that the system will not operate as expected.

In the field of programming, the remedy to this challenge lies in rigorous testing – repeated for so long and so meticulously that one obtains a high level of confidence that the program will exhibit predictable behavior in the most typical scenarios. Although many AI algorithms undergo testing and validation before

they are marketed, errors do sometimes surface once such systems are already in use. Examples include the image-recognition algorithm used in city surveillance cameras in the United States. It contained errors that resulted in the misidentification of certain individuals as suspects and misled the police. On top of this, many AI systems are experimental, and they are released without undergoing extensive tests that consider unusual scenarios. Risks associated with their use materialize in statistically very rare cases, but with time their likelihood increases. Consequently, we may witness such highly unlikely events as Shady's haywire expedition: under special circumstances, the drone accomplished a feat that its designers regarded as technologically impossible and may have created unexpected risks to air traffic, critical infrastructure, and more.

Errors

Another problem involving a lack of transparency in the processes adopted by AI systems lies in algorithms. Artificial intelligence is only as good as its algorithms. And these algorithms are created by humans, so they may be flawed and susceptible to errors. On the one hand, how an algorithm work results from limitations arising from machine design and the intricacies of their programming (constant patches, reedits, and last-minute changes made without the verification of their impact on overall performance). In addition, every algorithm is essentially a record of the expert knowledge that went into creating it. Algorithms could be described as rules developed on the basis of current knowledge on a specific topic and written down in the form of executable code – serving as instructions for machines or humans (such as a script for a job interview). We therefore have two sources of risk.

> Due to unintended AI bias, entire social groups could end up being excluded from access to certain services or resources.

One lies in the validity and relevance of the expert knowledge we use. Are the evaluation criteria employed by the experts we consult really appropriate for a given situation? The other involves the level of understanding of these principles by coders and their ability to translate this knowledge into instructions (code). Even the best knowledge can be implemented incorrectly, leading to the distortion of the intentions of the experts whose knowledge was leveraged. This can lead to AI bias. It occurs when AI systems are designed or trained in a way that results in unfair or discriminatory results. This poses a problem because AI is increasingly likely to be used to make decisions that affect people's lives, such as those concerning employee recruitment and credit ratings.

If these decisions are biased, they may entail significant negative consequences for individuals or groups. A faulty implementation of a perfectly reasonable rule may lead to entire groups of people being excluded from access to certain services or resources. Haste, technological limitations, and attempts to cut programming costs can result in the rules proposed by experts being overly simplified. The consequences may be very unpleasant, even appalling. Examples include an image recognition algorithm that repeatedly mislabeled black-skinned individuals as apes. Wanting simpler and shorter code, the coders had omitted the possibility of people having other skin colors than white. In this case, the absence of transparency and the failure to understand how the algorithm works can lead to egregious discrimination and flagrant injustice. In this context, a responsible and ethical approach to artificial intelligence necessitates developing standards of transparency and testing algorithms to uncover potential side effects.

Military applications

The third problem lies in the danger of using artificial intelligence in the military sector to perform combat-related tasks. Likewise, AI can be used for espionage and information manipulation, which could lead to serious problems in the area of national security. In this area, though, unexpected incidents may entail a lot more serious consequences than, for example, an intelligent washing machine that suddenly "goes rogue." We could ask: what if Shady had been an automated bomber carrying nuclear weapons? What if such a wayward machine decided to "come back home" and interpreted potential attempts to force it to stop as an enemy attack? War is difficult for machines to understand. In the case of AI systems capable of making autonomous decisions, military AI systems may be developed faster than their our capacity to keep them under control, which may lead to dangerous situations for humans and the environment. This is especially true for incidents in which AI systems will not be able to understand the context of their actions and will make decisions based on simple assumptions, instead of considering social, economic, and ethical aspects. During World War II, the Allies – despite boasting their superior ethics – made conscious decisions to demolish cities, including Dresden, which was of no strategic importance at the end of the war. They also decided to unleash nuclear weapons

ALEX YUZHAKOV/SHUTTERSTOCK.COM

How prepared is society for the potential dangers of artificial intelligence being used in the military sector, to perform combat-related tasks?

against Japanese cities, although they realized that Japan's capitulation was essentially a foregone conclusion. Another question is the issue of recognizing who the enemy is. How do we tell friend from foe? Even humans struggle with this challenge, often leading to incidents of "friendly fire" between units of the same forces. How will intelligent weapons know that the war is over? Automated, intelligent naval mines can not only strike ships that they misidentify as enemy vessels, but also avoid being detected and destroyed by countermeasure vessels, which may pose a threat to humans for decades after the end of an armed conflict.

Security

The fourth problem lies in AI system security and the risk that they may be hacked. Just like all computer systems, AI systems are vulnerable to hacking attacks. If an AI system becomes compromised, it can be used for malicious purposes. The more complex a system is, the more difficult it is to protect it from outside attacks. Unrecognized loopholes in algorithms also mean that a specific system can be misled. Examples include software for the identification of lung lesions. Researchers modified X-ray images by inserting images of the head of a gorilla. The algorithm saw them as alarming lesions in a critical area of the lungs. But people who analyzed the same images easily identified them as a suspicious artifacts suggesting that the data had been tampered with. What happens if a prankster hacks into a database of X-ray images and alters them for fun? How many patients could get misdiagnosed? What happens if an unauthorized person, either accidentally or on purpose, enters incorrect target coordinates into a combat drone?

To address security concerns, researchers and policymakers are working on AI security measures. One approach involves using the concept of "security by design." In other words, security should be a crucial issue firmly held in mind at every stage of the AI development process, from design to implementation.

Another approach is to build AI systems that are transparent and explainable – designed in a way that allows humans to understand how they work and make decisions. This can help reduce the risk of rogue AIs and AI bias.

> Under special, unforeseen circumstances, another AI-controlled military drone could end up "going rogue."

Yet another approach involves regulating AI development and implementation, including by establishing regulatory bodies to oversee these processes and laying down security standards and guidelines for AI developers to adhere to.

Artificial intelligence holds great potential to change the world for the better. As with all powerful technologies, however, it gives rise to security concerns that must be addressed. AI bias, dishonesty, and hacking attempts are just some of them. To ensure that AI is safe for humans, researchers and policymakers must work together to develop security measures that mitigate these risks. Only by doing so can we unlock the full potential of AI and shape a better future for everyone. ■

The author was helped by artificial intelligence, chiefly the web search engine Microsoft Bing, in gathering data for this article.

Further reading:

Bostrom, Nick (2014): *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Brockman, John (2019): *Possible Minds: Twenty-Five Ways of Looking at AI*. Penguin Press.

Tegmark, Max (2017): Life 3.0: *Being Human in the Age of Artificial Intelligence*. Penguin Press.