Capgemini
RESEARCH INSTITUTE

Conversations for tomorrow

Quarterly review
N°3 — 2021

Intelligent Industry: The Next Era of Transformation

#GetTheFutureYouWant

Marta Kwiatkowska
Professor of Computing Systems,
University of Oxford

BUILDING SAFER AI FOR THE NEXT ERA OF TRANSFORMATION

Marta Kwiatkowska is Professor of Computing Systems and Fellow of Trinity College, University of Oxford. She is known for fundamental contributions to the theory and practice of automated techniques for verification and correct-by-construction synthesis of systems from quantitative specifications, which have been adopted in diverse fields, including security, robotics, healthcare, DNA computing and nanotechnology. Her current research focus is on safety, robustness and fairness of AI decision making software.

Kwiatkowska is the first female winner of the Royal Society Milner Award, winner of the BCS Lovelace Medal and Van Wijngaarden award, and received an honorary doctorate from KTH Royal Institute of Technology in Stockholm. She is a member of the GPAI Working Group on Responsible AI, Fellow of the Royal Society, Fellow of ACM and Member of Academia Europea.

Artificial intelligence (AI) plays a key role in modern society. It drives cars, detects images, understands natural language, and controls complex industrial machines. When compared with traditional human-controlled operations, AI tends to be more consistent. In the near future, AI applications will take on greater autonomy in military, engineering, and industrial applications. However, these decision-making systems have critical exploitable flaws, which, if not addressed, will inevitably lead to loss of economic benefits, human life and, ultimately, trust in the technology.

The underlying method for building these AI systems is called deep neural networks (DNN). Loosely based on the neural networks in a human brain, they are vast and complex, but mathematically decipherable by normal human understanding. However, while mathematically transparent, logically they are “black boxes”: they work, but we don’t know how. If operators fail to remain vigilant, this fissure in our understanding of AI can expose it to adversarial exploitation.

Breaking an AI system

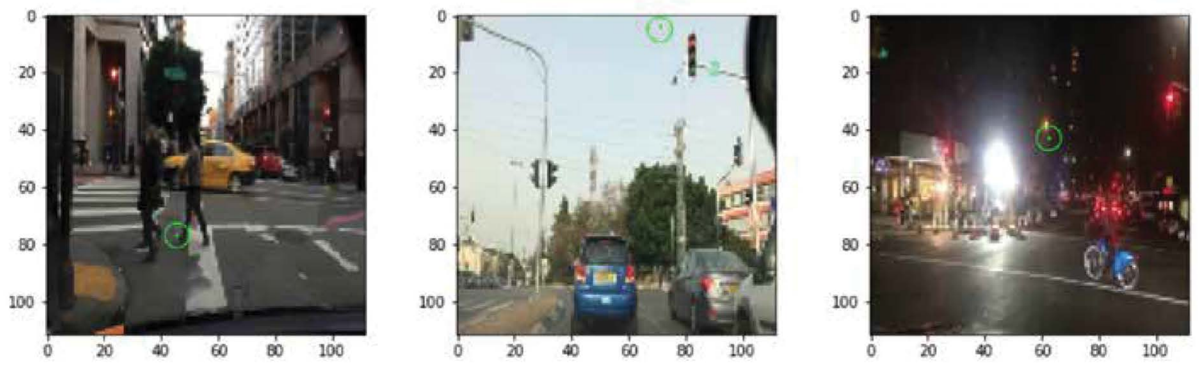
Research has shown that very simple changes can drastically impact an AI model’s outcomes, with potentially catastrophic consequences. Adversarial techniques[1] can fool the AI into misclassifying the input, even when the perturbation is minor. The Nexar Deep Learning Traffic Light Challenge, for example, has a database of 18,000 dashboard-camera images, to which the public has access and can contribute, to build AI models for traffic-light identification. The challenge is for researchers

> “Our research shows that the modification of just a few pixels can alter the AI object identification process completely – meaning a traffic light can be perceived as completely different object."

[1] Adversarial examples are inputs to machine-learning models designed to cause the model to commit a mistake.

to use technology that can identify and label each image as either "red," "green," or "null" (meaning no light has been detected). However, it requires only one inconsistent pixel to misguide the AI model into misclassifying the image in question, meaning that a red light can be recorded as green, or vice versa. "Moreover, these false classifications are often made with a high degree of confidence that they are correct (sometimes as much as 95%)."

Figure 1: A single pixel can drastically change outcomes, even using a state-of-the-art AI system



Source: Nexar Deep Learning Traffic Light Challenge. (a) Red light classified as green with 68% confidence after one pixel is changed. (b) Red light classified as green with 95% confidence after one pixel is changed. (c) Red light classified as green with 78% confidence after one pixel is changed.

These flaws have hugely significant change outcomes in computer-vision applications. A single stray pixel can easily overwhelm even a state-of-the-art vehicle-mounted AI system. Moreover, these adversarial examples are transferable, in the sense that an example misclassified by one network is also misclassified by a network with another architecture, even if it is trained on different data.

Implications of adversarial outcomes resonate across sectors

These simple input manipulations can cause large deviations of standard outcomes in autonomous cars. It could cause cars to drive into barriers, jump signals, or drive off road. While my group's research has shown this to be the case for cars, it can be easily applied to any image-identification use case from optical character recognition (OCR), handwriting interpretation, or natural language processing (NLP) systems.

A few other applications that can lead to adversarial outcomes are listed below:

1. **Natural language processing:**
   Today, natural language processing (NLP) software is regularly used to interpret legal documents and contracts.[2] These documents could be purposely designed to deliver flawed interpretation or impede progress. Similarly, this can be applied to language translation, speech-to-text applications, or document processing.

> **"In the near future, AI applications will take on greater autonomy in military, engineering, and industrial applications. However, if critical exploitable flaws in these decision-making systems are not addressed, they will inevitably lead to loss of economic benefits, human life and, ultimately, trust in the technology."**

[2] Thomson Reuters Blogs, "Legal AI: A beginner's guide," February 2017.

2. **Computer vision:**
   Our research shows that the modification of just a few pixels can alter the AI object identification process completely – meaning a traffic light can be perceived as completely different object. Applications range from remote sensing to radar systems and industrial quality control. With computer vision applications being the most successful and most critical application area, flaws exploited here can lead to suboptimal outcomes, economic loss and, in a worst-case scenario, even the loss of human life.

3. **Decision-making process:**
   Most decision-making systems utilize an array of inputs, from sensor-based or monitoring systems. More complex decisions are usually based on precedent. For example, if different sensors give different results, the critical decision making is based on prior probability outcomes. This means that digital applications such as finance and trading, cybersecurity, and healthcare can easily be intercepted through a critical input network.

Building safer AI systems

Building safer AI systems is the most critical challenge we are faced with today. Capgemini Research Institute's research into Ethics in AI shows that 60% of organizations have attracted legal scrutiny and 22% have faced a customer backlash in the last 2–3 years, owing to decisions reached by their AI systems.[3] The consequences for safety-critical systems will be a more drastic erosion of trust.

While a considerable research effort has gone into building more explainable, transparent, and robust AI systems, organizations and regulators can also take initial steps to mitigate these challenges:

1. **Foster awareness and understanding of possible adversarial exploitation**
   AI developers and teams usually have a singular focus on improving confidence rates and overall outcomes. This was the right direction to take when AI was in its infancy, as it helped establish AI as a tool that could be consistently useful to industry. However, with AI now being actively deployed in safety-critical systems, AI developers and teams need to understand the shortcomings of this traditional approach in building

[3] Capgemini Research Institute, "AI and the ethical conundrum," September 2020.

“We, at Oxford, are actively developing software tools to verify safety of AI systems, including diagnostic testing for the robustness issue relating to computer-vision applications. "

models, architectures, and autonomous decision-making systems. A more robust, safety-first approach is required.

2. **Develop tool chains to reduce exploitable flaws**
   It is well known that testing can detect software flaws but not prove their absence. A widely adopted method that can prove the correctness of software systems is model checking (an automated software technology to verify that given requirements are met for a variety of real-time embedded and safety-critical systems). Model checking techniques are today deployed by organizations such as Microsoft, Intel, and Facebook to check the correctness of their software. Model-checking methods for neural networks are still poorly understood, however; the development has been hampered by a lack of understanding of the theoretical fundamentals of neural networks, alongside their technical complexity. We, at Oxford, are actively developing software tools to verify safety of AI systems, including diagnostic testing for the robustness issue relating to computer-vision applications.[4]

[4] arXiv, “Feature-Guided Black-Box Safety Testing of Deep Neural Networks,” Matthew Wicker, Xiaowei Huang, Marta Kwiatkowska, In Proc. 24th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS’18).

3. **Regulators need to build safety guidelines and testing frameworks for safety-critical AI systems**
   Regulators also need to put emphasis on developing robustness criteria for safety-critical AI systems and frameworks for checking that such criteria are met. Standardized testing and evaluation frameworks should be created to support the development of safety-critical autonomous systems. These should extend the existing safety regulations found for cars, medical devices, and the workplace.

4. **Develop collaborative research into AI systems and their associated transparency, and ethical status**
   The current field of adversarial exploitation and model checking for neural networks is still in its infancy and we still have a long way to go to establish a complete understanding of it. Industry-wide collaboration is required to guide the development of appropriate frameworks and standards and to develop new ways of working. Collaboration is required to build open-source tool chains and evaluation methodologies and to govern practices among AI developers and teams.

Adversarial AI is still in its infancy in terms of industry understanding. To date, there has been no (detected) concerted effort to exploit these loopholes. However, it is only a matter of time before hostile players work to exploit them. Currently, as well as the potential involvement of hostile actors, these AI systems also show potential flaws relating to a sensitivity to naturally occurring "noise" in the environment. The reliability, robustness, and possible economic value of AI is directly linked to the trust we have in these systems. A significant effort to address these challenges is required to ensure we fulfil the social and economic potential of AI.

www.capgemini.com