CENTRE FOR THE STUDY OF EXISTENTIAL RISK

UNIVERSITY OF CAMBRIDGE

Response to the European Commission's High-Level Expert Group on Artificial Intelligence
Draft Ethics Guidelines for Trustworthy AI
Working Document for stakeholders' consultation

We are writing from the Centre for the Study of Existential Risk, a research group at the University of Cambridge which studies the security implications of emerging technologies. For the last five years we have been closely involved with the European and international debate about the ethical and societal implications of artificial intelligence (AI).

These Draft Ethics Guidelines are an important, concrete step forward in the international debate on AI ethics. In particular the list of technical and non-technical methods and the assessment list will be useful to researchers and technology company employees who want to ensure that the AI systems they are busy developing and deploying are trustworthy.

"The list of "Requirements of Trustworthy AI" is a useful one. 'Robustness' and 'Safety' are particularly important requirements. They are both often individually mentioned in sets of AI principles, and there are extensive and distinct fields of study for each of them. Robustness is an important requirement because our AI systems must be secure and able to cope with errors. Safety is an important requirement as our AI systems must not harm users, resources or the environment.

Robustness and safety are crucial requirements for trustworthiness. As an analogy, consider that we could not call a bridge 'trustworthy' if it was not reliable and resilient to attack, and also safe for its users and the environment. These two requirements are importantly distinct from the other requirements, and work best as stand-alone requirements."

***

The report "invite[s] stakeholders partaking in the consultation of the Draft Guidelines to share their thoughts on additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI."

We would like to share some additional technical and non-technical methods that are not yet on the list. These are mostly drawn from the major Febuary 2018 report *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* We co-authored this report with 26 international experts from academia and industry to assess how criminals, terrorists and rogue states could maliciously use AI over the next five years, and how these misuses might be prevented and mitigated.

When released this report was covered across Europe and welcomed by experts in different domains, such as AI policy, cybersecurity, and machine learning. We have

subsequently consulted several European governments, companies and civil society groups on the recommendations of this report.

The European Union's Coordinated Plan on Artificial Intelligence, published on the 7th of December 2018, mentions the importance of the security-related AI applications and preventing malicious use:

"2.7. Security-related aspects of AI applications and infrastructure, and international security agenda: There is a need to better understand how AI can impact security in three dimensions: how AI could enhance the objectives of the security sector; how AI technologies can be protected from attacks; and how to address any potential abuse of AI for malicious purposes."

Several of the methods we explored are already mentioned in the Guidelines, such as codes of conduct, education and societal dialogue. However we also explored some methods that you do not yet mention. Our report made recommendations in four 'priority research areas'. In this response we split these into 'technical' and 'non-technical' methods.

- Learning from and with the Cybersecurity Community
- Exploring Different Openness Models
- Promoting a Culture of Responsibility
- Developing Technological and Policy Solutions

Technical methods include:

Learning from and with the Cybersecurity Community

Formal verification. The use of mathematical methods to offer formal proofs that a system will operate as intended. In recent years this has worked on complex systems, including the CompCert compiler and the seL4 microkernel. It could be applied to AI systems.

Security tools. Software development and deployment tools now include an array of security-related capabilities (testing, fuzzing, anomaly detection, etc.). Tools could be developed to make it standard to test and improve the security of AI components during development and deployment. Tools could include: automatic generation of adversarial data; tools for analysing classification errors; automatic detection of attempts at remote model extraction or remote vulnerability scanning; and automatic suggestions for improving model robustness.

Secure hardware. Increasingly, AI systems are trained and run on hardware that is semi-specialized (e.g. GPUs) or fully specialized (e.g. TPUs). Security features could be incorporated into AI-specific hardware to, for example, prevent copying, restrict access, and facilitate activity audits.

Exploring Different Openness Models

Central access licensing models. In this emerging commercial structure, customers use services (like sentiment analysis or image recognition) from a central provider without having access to the technical details of the system. This model could provide widespread use of a given capability while reducing malicious use by, for example: limiting the speed of use, preventing some large-scale harmful applications; and explicitly prohibiting malicious use in the terms and conditions, allowing clear legal recourse.

Promoting a Culture of Responsibility

Differentially private machine learning algorithms. These combine their training data with noise to maintain privacy while minimizing effects on performance. There is increasing research on this technological tool for preserving user data privacy.

Secure multi-party computation. MPC refers to protocols that allow multiple parties to jointly compute functions, while keeping each party's input to the function private. This makes it possible to train machine learning systems on sensitive data without significantly compromising privacy. For example, medical researchers could train a system on confidential patient records by engaging in an MPC protocol with the hospital that possesses them.

Coordinated use of AI for public-good security. AI-based defensive security measures could be developed and distributed widely to nudge the offense-defense balance in the direction of defense. For example, AI systems could be used to refactor existing code bases or new software to security best practices.

Monitoring of AI-relevant resources. Monitoring regimes are well-established in the context of other dual-use technologies, most notably the monitoring of fissile materials and chemical production facilities. Under certain circumstances it might be feasible and appropriate to monitor inputs to AI technologies such as hardware, talent, code, and data.

Non-technical methods include:

Learning from and with the Cybersecurity Community

Red teaming. A common tool in cybersecurity and military practice, where a "red team" composed of security experts deliberately plans and carries out attacks against the systems and practices of the organization (with some limitations to prevent lasting damage), with an optional "blue team" responding to these attacks. Extensive use of red teaming to discover and fix potential security vulnerabilities and safety issues could be a priority of AI developers, especially in critical systems.

Responsible disclosure of AI vulnerabilities. In the cybersecurity community, "0-days" are software vulnerabilities that have not been made publicly known, so defenders have "zero days" to prepare for an attack making use of them. It is common practice to disclose these vulnerabilities to affected parties before publishing widely about them, in order to provide an opportunity for a patch to be

developed. AI-specific procedures could be established for confidential reporting of security vulnerabilities, potential adversarial inputs, and other types of exploits discovered in AI systems.

Forecasting security-relevant capabilities. "White-hat" (or socially-minded) efforts to predict how AI advances will enable more effective cyberattacks could allow for more effective preparations by defenders. More rigorous tracking of AI progress and proliferation would also help defensive preparations.

Exploring Different Openness Models

Pre-publication risk assessment in technical areas of special concern. In other dual-use areas, such as biotechnology and computer security, the norm is to analyse the particular risks (or lack thereof) of a particular capability if it became widely available, and decide on that basis whether, and to what extent, to publish it. AI developers could carry out some kind of risk assessment to determine what level of openness is appropriate for some types of AI research results, such as work specifically related to digital security, adversarial machine learning, or critical systems.

Sharing regimes that favour safety and security. Companies currently share information about cyber-attacks amongst themselves through Information Sharing and Analysis Centers (ISACs) and Information Sharing and Analysis Organizations (ISAOs). Analogous arrangements could be made for some types of AI research results to be selectively shared among a predetermined set of 'trusted parties' that meet certain criteria, such as effective information security and adherence to ethical norms. For example, certain forms of offensive cybersecurity research that leverage AI could be shared between trusted organizations for vulnerability discovery purposes, but would be harmful if more widely distributed.

Promoting a Culture of Responsibility

Whistleblowing measures. Whistleblowing is when an employee passes on potentially concerning information to an outside source. Whistleblowing protections might useful in preventing AI-related misuse risks.

Nuanced narratives. There should be nuanced, succinct and compelling narratives of AI research and its impacts that balance optimism about its vast potential with a level-headed recognition of its challenges. Existing narratives like the dystopian "robot apocalypse" trope and the utopian "automation boon" trope both have obvious shortcomings. A narrative like "dual-use" might be more productive.