AI ANXIETIES IN THE CHEMICAL AND BIOLOGICAL WEAPONS PROHIBITION REGIMES

Artificial intelligence (AI) has the potential to revolutionize scientific, economic, and social activities across diverse sectors. However, there is growing concern about the negative impacts that AI's dual-use applications and outcomes may bring. These "AI-anxieties" are both real and imagined, rooted in what is possible today and what may be possible in the future.

Concern about the potential dual-use applications of AI have also been expressed within the context of chemical and biological weapons (CBW) prohibition regimes, where AI's capacity to enable the design, development, deployment, and detection of CBW is starting to be considered.

Emerging narratives that suggest AI could facilitate the development of super toxic agents, or provide low-cost routes for state and non-state actors to develop and employ CBW, are capturing public imagination.

This briefing note presents four emerging AI-anxieties concerning CBW, using existing ethical principles for responsible AI as a guide to thinking about how our norms and values can help address these challenges.

By identifying what characteristics of AI these principles seek to mitigate, we can better visualize how those AI characteristics might generate specific challenges within the CBW prohibition regimes. This is an early-stage approach that can contribute to emerging efforts to understand the nature of these emerging challenges.

Key Points

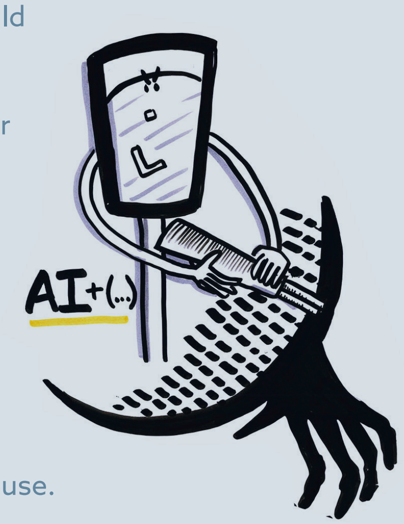
- AI is not a stand-alone entity, but a component within a larger system that combines with other technologies to enhance data processing and decision making.
- While AI can introduce new challenges and anxieties, many of these are not fundamentally different from those associated with other technologies or practices. AI can amplify and modify existing challenges, and must be viewed in the broader context of the systems in which it operates.
- Mitigating the potentially harmful effects of AI does not require entirely new sets of governance architectures but rather, by identifying and addressing potential AI-anxieties, can help to inform the amendment and augmenting of existing frameworks.

UNIVERSITY OF SUSSEX

HSP THE HARVARD SUSSEX PROGRAM

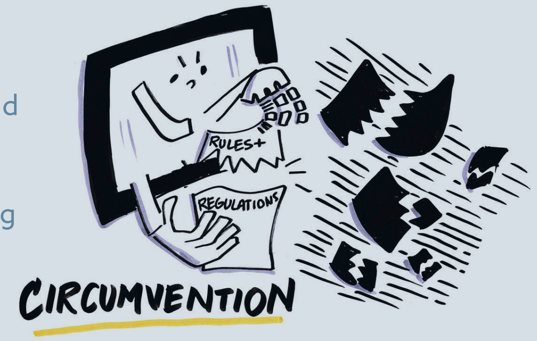
AI + (TECHNOLOGY)

The integration of AI with other technologies poses significant challenges to the CBW prohibition regimes. Scenarios conforming to this anxiety revolve around the ways in which AI could augment or amplify the risks from existing dual-use technologies. For example, use of AI targeting in autonomous vehicles for delivery of riot control agents (RCAs) may deepen ambiguity around the safe and legal employment of RCAs under the CWC. Alternatively, the coupling of AI with advanced robotics could enable actors with limited human resource to design and synthesise chemicals or biologics remotely. Ethical principles in this case may conform more to 'Safe and Secure' and 'Human-centric' values where the infrastructure, governance, and assessments of intent can help to clarify potential governance measures in AI use.



PATHWAYS TO CIRCUMVENTION

AI could help to overcome traditional limitations, obstacles and controls that hinder the development, storage, and deployment of CBWs. For instance, AI may have the potential to enhance efforts by malicious actors to circumvent import/export controls by identifying or modelling unlisted precursors for chemical synthesis. Additionally, the use of AI can reduce the number of human resources needed to develop or deploy these weapons, resulting in decreased costs and avoiding potential 'moral dilemmas.' Furthermore, AI could be utilized to design agents that are more difficult to detect, and may be harder to defend against. In these cases, we might expect values of 'Controllability' and 'Accountability' to reflect back a need for examining governance in AI development.



AI SHOULD BE...

Explainable and Accountable

"The ability to explain how and why particular outcomes were reached; and that the data inputs, design structure and operating systems, overall operation and outcomes, should permit accountability exercises."

Human-centric

"Humans are ultimately responsible for the full design and operational cycle of AI and its outputs, and understanding where and how humans are implicated can support assessments of intent."

Controllable

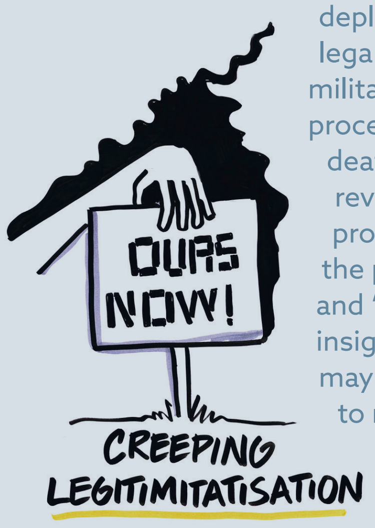
"Ability to control inputs, workings, and outcomes, so as to redirect, amend, override or shut down operations; outputs must also have potential for controllability (i.e. prevention of automatic dissemination)."

Safe and Secure

"AI requires cyber and infrastructural security to protect from malign actors; internally requires fallbacks, ability to stop/ override; review mechanisms to ensure outcomes do not present dangers to humans."

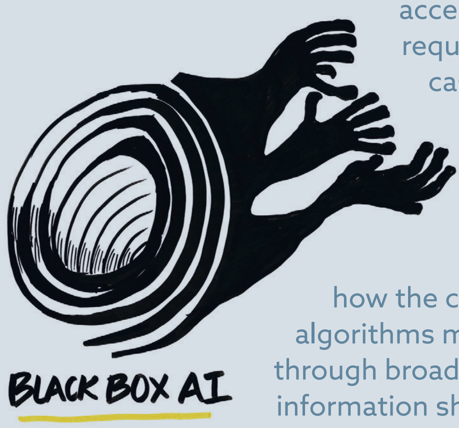
CREEPING LEGITIMISATION

The insidious erosion of legal prohibitions and social taboos through certain lines of research and development is not a new concern. However, AI's potential to facilitate, accelerate, and create new pathways for research that have small crossovers from legitimate to illegitimate may provide nefarious actors with new opportunities to develop agents and technologies that could challenge CBW treaties. For example, AI's exacerbation of existing assumptions may produce unintended or ambiguous deployment of RCAs that stretch legal boundaries. Moreover, military research into biological life processes toward 'war-without-death' outcomes, could be revitalised by the opportunities provided by AI technologies. Here, the principles of 'Accountability' and 'Human-centric' may offer an insight into how different actors may be governed (or may govern) to maintain legal and social boundaries in light of novel capabilities.



THE BLACK BOX

The opaque nature of AI's inputs, methods, and outputs creates additional transparency burdens for developers and users. This intangibility, opacity, and lack of explainability of AI systems will permeate into, and characterise, challenges in many areas. For example, in verifying allegations of use, tracing accountability may be complicated if an off-the-shelf AI tool is used with limited knowledge of the tool's exact provenance, training data, and methods of deduction. Alternatively, the use of AI tools to develop and produce potential CBWs may limit attribution in cases where multiple actors have access to the tools and data required to do so. In these cases, the principles of 'Explainable', 'Accountable', and 'Controllable' come to the fore in understanding how the creation and use of AI algorithms might be monitored through broader transparency and information sharing.



Implications

The four AI-anxieties demonstrate that integration of AI technologies do not necessarily produce entirely new categories of threat or risk, but that they embed within existing ones. Importantly, specific scenarios fit within multiple categories of anxiety depending on the ways in which AI is interacting with existing technologies and modes of governance. This recognition requires anxieties about AI to be grounded within the contexts and challenges that actors operating in the CBW prohibition regimes are already familiar with. Within those contexts, AI's impacts will likely be to:

- Accelerate research and development processes;
- Open up new prospective research pathways;
- Degrade transparency;
- Complicate information providence, relevance, and consequence.

The next stage is then to assess whether existing forms of governance are able, through existing capability or through modification, to mitigate the impact of AI on existing and emerging challenges.

The four ethical principles for responsible AI direct us to the sort of questions we should be asking in those efforts to visualise, characterise, and minimize the negative implications of AI technologies.

For further information:

This research was conducted by the Harvard-Sussex Program at the University of Sussex Business School's Science Policy Research Unit. The research project was kindly funded by the UK Foreign, Commonwealth, and Development Office's Counter Proliferation and Arms Control Centre. Our thanks go to Shaunna McIvor and Boaz Chan who provided excellent assistance on the project.

Principal Investigator: Dr. Joshua R Moon
E-mail: J.R.Moon@sussex.ac.uk

How can AI be better governed in CBW contexts?

1. Contextualize AI development and applications to better understand policy-specific challenges.
2. Socialize AI by identifying actors involved in development and purpose of use.
3. Examine AI-anxieties in light of how they may impede or weaken current efforts in CBW prohibition.
4. Evaluate existing governance actors and architectures to understand where opportunities and gaps exist to address these AI-anxieties, and, where possible, rectify them.
5. Focus on existing tools and mechanisms to address challenge-specific AI anxieties.

Next steps for better AI policy:

- Building more detailed potential scenarios of AI's impact on CBW prohibitions, specifically providing detail on the use of technologies and the roles of different actors within those scenarios, will enable better understandings of how risks can be lessened;
- Understanding the impacts and consequences of AI in context and in practice will enable deeper knowledge of how AI might be governed to preserve its positive impacts on society;
- Mapping and evaluating existing governance for CBW prohibition regimes with specific reference to detailed scenarios will provide policymakers with a better understanding of where AI interacts with different webs of prevention and how those webs can be developed to accommodate this new area of science and technology.