HYBRID COE

Hybrid CoE Strategic Analysis / 26

MAY 2021

Cyber-biosecurity: How to protect biotechnology from adversarial AI attacks

ELEONORE PAUWELS

Hybrid CoE

Hybrid CoE Strategic Analysis is typically a short paper written by academic and research community experts. Strategic Analyses are based on long-term research experience, or on current or completed research projects. The Strategic Analysis papers enhance understanding of different phenomena in the realm of hybrid threats. They do not offer direct recommendations but aim to explain processes and identify gaps in knowledge and understanding, as well as highlight trends and future challenges. Each Strategic Analysis paper includes a bibliography for further reading. Topics are related to Hybrid CoE's work in all of its main functions.

The European Centre of Excellence for Countering Hybrid Threats tel. +358 400 253800 www.hybridcoe.fi

ISBN (web) 978-952-7282-72-4
ISBN (print) 978-952-7282-73-1
ISSN 2670-2282

May 2021

Hybrid CoE is an international hub for practitioners and experts, building Participating States' and institutions' capabilities and enhancing EU-NATO cooperation in countering hybrid threats, located in Helsinki, Finland.

The responsibility for the views expressed ultimately rests with the authors.

Cyber-biosecurity: How to protect biotechnology from adversarial AI attacks

> *The digital infrastructure that underpins biotechnology is a global public good – and a growing target for data manipulation and adversarial information operations. Emergent hybrid threats that compromise AI- and cyber-security within the bio-economy are contributing to a new geopolitics of inequality and insecurity that cuts across societies and borders. Protecting information integrity, explainability, and public trust in modern biotechnology is becoming a substantial asset to preserve both global security and national sovereignty* – writes Eleonore Pauwels, international expert on converging technologies, and Senior Fellow with the Global Center on Cooperative Security.

In the last two decades, **the world of biotechnology has moved from analogue to digital, converging with artificial intelligence (AI) as an innovation catalyst**. New collaborations between AI, geneticists and bio-engineers have led to the field of functional genomics, a more precise understanding and optimization of functional processes in genome biology. Deep-learning algorithms can help analyze and test genetic functions in silico, and help predict the effect of a genetic mutation on an individual's overall genome. Such algorithms improve analysis of the combinatorial relationship between genotype and phenotype in genomic datasets related to humans and pathogens. Other deep-learning models aim to unveil important features of genome biology, from simulating RNA-processing events to modelling the genetic regulatory code governing gene expression.

The new frontier of functional genomics is therefore increasingly happening "in silico", producing important knowledge insights that build on synthetic datasets as well as algorithmic and advanced computing. Substantial progress will also derive from digitizing, processing and learning from genomics and other multimodal omics datasets that are part of comprehensive approaches to analyzing complete genetic or molecular profiles of humans and pathogens**. Functional genomics, and biosciences in general, are becoming not only crucial and sensitive digital assets, but also critical information infrastructure.** Transformational opportunities range from improving trust in precision medicine diagnoses and therapies, to ensuring reproducibility and efficiency in complex biotech supply chains, and isolating potential harmful genetic functions in biosecurity screening.

The integration of biotechnology with AI is emerging as a geo-strategic, societal, and welfare asset that can define a country's digital sovereignty and preserve national and international security. In the absence of a robust AI and cybersecurity framework, however, **AI can be misused to manipulate datasets in seconds, creating hybrid insecurity flashpoints** and leading to widespread collective data harms, research and industrial sabotage, as well as compromised governance systems and data integrity crucial to health, food and civilian security.

This Strategic Analysis aims at uncovering a new typology of AI-led cyberthreats that can manipulate and corrupt the integrity of datasets and algorithmic models crucial to the global knowledge-production cycle in bio-medicine, biotechnology and biosecurity. It demonstrates how **such emergent hybrid threats may not only produce lethal outcomes for populations and erode countries' digital sovereignty, but also drastically undermine public trust in the bio-economy's critical information and governance infrastructures**.

Emergent hybrid and converging threats

Hacks of biomedical and genomic datasets have already resulted in the manipulation of sensitive information, from cancer data in patients' CT scans to the DNA sequences of individuals' genomes. In December 2020, IBM's threat intelligence task force exposed a global phishing campaign targeting organizations associated with a COVID-19 cold chain.[1]

These dynamics mark a powerful socio-technical shift influencing how the convergence of AI with cybersecurity and biotechnology is changing regional security landscapes and posing complex transnational challenges requiring multilateral responses. **Targeting biomedical datasets and the digital infrastructure of the bio-economy is increasingly being used by state and non-state actors alike for adversarial or commodification purposes, with the potential to sabotage or weaponize biomedical research, biotech facilities, and biomanufacturing supply chains**.

Motivations behind adversarial attacks on the biotechnology sector range from falsifying clinical trials and research, holding the integrity of biomedical data hostage, undermining trust in precision medicine diagnoses and treatments, and sabotaging critical infrastructure for health, food and bio-security. Most AI security studies have emphasized how adversarial attacks are easy to engineer and do not require outstanding technical expertise. Moreover, they are hard to detect and can transfer to many bio-computing domains.

Manipulating biomedical research and population datasets

The integration of AI computing within modern biomedicine allows researchers to rely on **synthetic datasets and predictive methods** to produce actionable knowledge in a genome's biology and assess its clinical value. Such new digital infrastructure is an asset for research and knowledge production, **with implications not only for precision medicine and its clinical application, but also for infectious disease prevention and control, as well as effective medical countermeasures and management of public health crises**.

The digital interdependence of modern biosciences subjects our growing functional intelligence about genome biology to new information risks, particularly **adversarial attacks that could corrupt the integrity of biological datasets and manipulate the functioning of deep-learning analysis systems**. Several studies in AI security have demonstrated how generative adversarial networks can be trained to drastically undermine the predictive ability of a wide range of medical image analysis systems that are based on deep learning. In 2018, researchers at Ben-Gurion University designed a malicious attack to manipulate cancer data in hospital CT scans, generating false lung tumours that conformed to a patient's unique anatomy, leading to a misdiagnosis rate in excess of 90%.[2] Furthermore, researchers at Harvard University tested adversarial attacks against algorithms used to diagnose skin cancer images, showing that such attacks only required modifying a few pixels in the original biopsy picture to corrupt a diagnosis.[3] As medical intelligence about the treatment of cancers, blood clots, brain lesions and infections could be manipulated, **adversarial attacks on deep learning pose a substantial risk to our most critical medical and clinical infrastructures**.

The attack surface extends far beyond medical diagnosis and clinical trials with **adversarial malware that could target the integrity of genomics and other omics datasets related to humans and pathogens**. Researchers at Sandia National Laboratory have demonstrated how autonomous malware could be used to manipulate raw data within large curation of human genomes.[4] The malicious malware could be used to target the functioning of genetic analysis software and alter actual fragments of DNA sequences within individuals' genomes. Such malicious tampering could result in

[1] Zaboeva, 'IBM Uncovers Global Phishing Campaign Targeting the COVID-19 Vaccine Cold Chain'.
[2] Mirsky et al., 'CT-GAN: Malicious Tampering of 3D Medical Imagery Using Deep Learning'.
[3] Finlayson et al., 'Adversarial Attacks on Medical Machine Learning', 1287-1289.
[4] Sandia National Laboratories, 'Personalized medicine software vulnerability uncovered by Sandia researchers'.

misdiagnosis with an impact on clinical decisions. **This type of data poisoning could affect in silico predictive models in functional genomics, including how we diagnose and treat complex genetic diseases, how we analyze and study the pathogenicity of viral and microbial threats, and how we develop adequate medical countermeasures** for subgroups of patients. What is at stake is the global knowledge-production cycle in biomedicine.

Sabotaging bio-engineering and bio-manufacturing

New capacities in automation and remote manufacturing – including cloud laboratories and commercial DNA sequencing and synthesis – are accelerating the decentralization of bio-engineering experiments. Increasingly, **biotech laboratories and bio-pharmaceutical manufacturing systems are automated, equipped with AI analytics software and connected to cloud services. On such platforms, technical skills and tacit knowledge are encoded into "automated protocols" that program and standardize the instructions of a biotech experiment.** Equipped with connected sensors to measure experimental variables, the AI operating system uses constant learning and iteration to augment the precision of automated protocols and may even lead to the in silico design of novel experiments with less outside guidance. Automated laboratories therefore offer advantages that are crucial to precision medicine as they allow for scalability, reproducibility and outsourcing to a broader and more diverse talent pool.

The advent of autonomy provides an increasing potential to weaponize biotech laboratories and biomanufacturing supply chains through adversarial attacks waged in cyber-operations. Adversarial algorithms could target vulnerabilities in automated protocols to corrupt networks of sensors and duly impact control decisions related to important experimental parameters. **Resulting harm could range from producing pharmaceutical products that do not match specification standards (leading to waste and shortage), to spoiling vital stocks of vaccines, antibiotics, cell or immune therapies for cancer treatment.** Cyber criminals and state actors have already mounted targeted cyber-operations against firms researching, producing and distributing COVID-19 vaccines. In December 2020, IBM researchers and the US Cybersecurity and Infrastructure Security Agency (CISA) unveiled global social engineering attacks "intended to steal the network log-in credentials of corporate executives and officials at global organizations involved in the refrigeration process necessary to protect vaccine doses".[5] The underlying goal could be to access and manipulate shared information about how the vaccine is shipped, stored, kept cold and delivered.

Weaponizing biotech laboratories could escalate into a strictly biothreat-based scenario while avoiding traditional screening and oversight. Automated bio-labs could be used to 1) produce toxins that can disrupt cellular metabolism, 2) synthetize a known lethal pathogen, or 3) use gene-editing to augment the capacity of a pathogen to infect a host, evade the immune system, spread among subpopulations or resist vaccines or antibiotics. **An area of near-term concern is the automated design of bacteria with multidrug resistance or the modification of commensal bacteria to become super-producers of toxins.**

The convergence of AI and automation with biotechnology is increasingly challenging the compliance tools, verification methods and overall oversight that countries can rely on to ensure non-proliferation within the current disarmament regime, the Biological Weapons Convention (BWC). Importantly, **no adequate guidance exists to prevent the adversarial use of biological data and algorithmic models to produce pathogens of concern, or produce a biosecurity consequence by exploiting vulnerabilities in the cyber-biosecurity infrastructure.**

Hacking/corrupting biosecurity screening

By improving our knowledge of DNA functions, AI computing is becoming an integral part of biosecurity screening mechanisms. In particular,

[5] Sanger & LaFraniere, 'Cyberattacks Discovered on Vaccine Distribution Operations'.

algorithmic models are instrumental in preventing illicit gene synthesis and illicit experiments in gain-of-function research, a field that studies the potential to enhance the transmissibility or pathogenicity of potential pandemic pathogens. Government-funded programmes are already designing deep-learning systems to predict how genetic sequences are meant to function, before being assembled, and even if the combination is not seen in nature. Gene-synthesis companies are developing computational threat models that can be applied to characterize the function of novel combinations of DNA sequences. Similar algorithmic tools play an increasing role in microbial forensics, using their capacity for anomaly detection to identify the specific signatures left in modified organisms.

Adversarial attacks could be designed to corrupt the predictive ability of screening algorithms to identify threats based on functional analysis of DNA sequences. By obfuscating functional data from sequences of pathogen and toxin DNA, generative adversarial networks could manipulate the integrity of the priority dataset shared by stakeholders to train screening algorithms. **Such data manipulation could drastically undermine the confidence level of screening algorithms when they aim to ascribe threat potential to known and unknown genes, including genes responsible for the pathogenesis of viral threats, bacterial threats and toxins**. Both human and algorithmic understanding of functional genetic data is still weak and fraught with complex unknowns. Adversarial attacks therefore have a very high potential to succeed in undermining stakeholders' trust in DNA screening.

Conclusions: Protecting the bio-economy

The digital infrastructure that supports biomedicine and biotechnology is a global public good. Emergent hybrid threats that compromise AI- and cyber-security within the bio-economy are contributing to a new geopolitics of inequality and insecurity that cuts across societies and borders.

Adversarial information operations that target the biotechnology sector are a powerful type of hybrid threat. They may serve an array of offensive goals and involve broad coalitions of malicious actors, including states, non-state actors and surrogates. **They target systemic vulnerabilities and different civilian and security interfaces, from population datasets and industry's clinical trials to biosecurity screening. They also interfere with diverse levels of strategic and emergency decision-making.**

New forms of covert, adversarial data manipulation are extremely hard to detect, creating new challenges for attribution. What is potentially under attack is the global knowledge-production cycle in biosciences. The aim is not only to seriously **erode a country's digital sovereignty, but also to undermine both global leadership crisis response and people's trust and resilience.**

Combinations of poisoning population datasets, falsifying biomedical research, sabotaging bio-manufacturing and corrupting biosecurity screening would have drastic economic costs and potentially lethal outcomes for populations. Yet **the most damaging impact would be on citizens' trust in governing institutions, emergency data systems, industrial laboratories, food supply chains, hospitals and critical health infrastructures.** This could have powerful, long-term implications for peace and security. As vulnerable states are unable to prevent and mitigate data-poisoning attacks, they could become fertile operating grounds for cyber mercenaries, terrorist groups and other actors, increasingly compromising the data integrity and robustness of our globalized intelligence system.

Policymakers need to start working with technologists to better understand the security risks emerging from the combination of AI and biotechnology, and the implications for the bio-economy and its critical information systems. Preventing and mitigating such threats requires a substantial departure from legacy approaches conceived to contain biological weapons by strictly controlling physical access to biotechnological equipment, materials, and listed bio-agents. First, policymakers and technologists should **use foresight to anticipate and more clearly determine the functional definitions of dual use that are emerging across AI, cyber and biosecurity domains.** Second, they should collaborate to identify and **assess the potential vulnerabilities that can be exploited in the convergence of AI, cyber- and biotechnologies to cause extensive civilian harm, produce biosafety and biosecurity incidents, and compromise the knowledge-production cycle of the bio-economy**. The new digital vulnerabilities emerging in biomedicine and biotechnology will increasingly require updated standards and practices, such as data-authentication and verification mechanisms that do not exist in our legacy policy, cyber- and bio-security frameworks.

Author

Eleonore Pauwels is an international expert on the security, societal and governance implications generated by the convergence of artificial intelligence with other dual-use technologies, including cybersecurity, genomics and genome editing, and Senior Fellow with the Global Center on Cooperative Security. She provides expertise to the World Bank and the United Nations, among others, and works closely with governments and private sector actors on AI-Cyber threat prevention, the changing nature of conflict, foresight and global security.

Bibliography

Allyn, Jérôme et al. 'Adversarial attack on deep learning-based dermatoscopic image recognition systems'. *Medicine*, vol. 99, no 50 (December 2020). PubMed Central, doi:10.1097/MD.0000000000023568.

Bipartisan Commission on Biodefense. 'Cyberbio Convergence: Characterizing the Multiplicative Threat', 17 September 2019. https://biodefensecommission.org/events/cyberbio-convergence-characterizing-the-multiplicative-threat/. Last accessed 28 April 2021.

Finlayson, Samuel G. et al. 'Adversarial Attacks on Medical Machine Learning'. *Science*, vol. 363, no 6433 (March 2019). PubMed, doi:10.1126/science.aaw4399.

Hudson, Corey M. 'From Buffer Overflowing Genomics Tools to Securing', DEF CON 27 Bio Hacking Village, DEFCON Conference. https://www.youtube.com/watch?v=7du1TltZOJg. Last accessed 28 April 2021.

Mirsky, Yisroel et al. 'CT-GAN: Malicious Tampering of 3D Medical Imagery Using Deep Learning', Cornell University, January 2019. https://arxiv.org/abs/1901.03597v3. Last accessed 28 April 2021.

Pauwels, Eleonore. 'From Drone Swarms to Modified E. Coli: Say Hello to a New Wave of Cyberattacks', World Economic Forum 2019. https://www.weforum.org/agenda/2019/05/cyber-attacks-ai-artificial-intelligence-drone-swarms-data-poisoning/. Last accessed 28 April 2021.

Pauwels, Eleonore. 'The New Geopolitics of Converging Risks', United Nations University 2019. https://collections.unu.edu/eserv/UNU:7308/PauwelsAIGeopolitics.pdf. Last accessed 28 April 2021.

Pauwels, Eleonore. 'What's needed to prevent cyberbiosecurity threats and protect vulnerable countries', World Economic Forum 2020. https://www.weforum.org/agenda/2020/06/prevent-cyber-bio-security-threats-covid19-governance/. Last accessed 28 April 2021.

Sandia National Laboratories. 'Personalized medicine software vulnerability uncovered by Sandia researchers', Sandia Labs News Releases, 1 July 2019. https://share-ng.sandia.gov/news/resources/news_releases/genomic_cybersecurity/. Last accessed 28 April 2021.

Sanger, David E. & LaFraniere, Sharon. 'Cyberattacks Discovered on Vaccine Distribution Operations', *The New York Times*, 3 December 2020. https://www.nytimes.com/2020/12/03/us/politics/vaccine-cyber-attacks.html. Last accessed 28 April 2021.

Segal, Michael. 'An Operating System for the Biology Lab'. *Nature*, vol. 573, no 7775 (September 2019): S112-13. doi:10.1038/d41586-019-02875-z.

Zaboeva, Claire. 'IBM Uncovers Global Phishing Campaign Targeting the COVID-19 Vaccine Cold Chain'. *Security Intelligence*, 3 December 2020. https://securityintelligence.com/posts/ibm-uncovers-global-phishing-covid-19-vaccine-cold-chain/. Last accessed 28 April 2021.

Hybrid CoE