UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Humanities
Fakulteit Geesteswetenskappe
Lefapha la Bomotho

Humanities 100
Since 1919

AI Policy and Ethical Issues in Industry

AI Ethics in Nuclear Security

Emma Ruttkamp-Bloem

8 February 2024

UNESCO

SACAIR
Southern African Conference for Artificial Intelligence Research

World Institute for Nuclear Security

The Role of Artificial Intelligence in Strengthening the Security of Nuclear Facilities

cair
CENTRE FOR ARTIFICIAL INTELLIGENCE RESEARCH

Outline

General AI Ethics review

Ethics of Nuclear in AI (ENAI)

The use of AI in Nuclear Security

- Pro’s
- Con’s

Way forward

The ethics of AI in nuclear security

Reflection

Why care about AI?

The Good:

Endless list from democratizing healthcare to revolutionizing food safety

The Bad:

There are concerns around e.g., the amplification of inequality, threats to social justice and political stability, the quality and integrity of information, privacy and the right to mental integrity, and threats to the environment and ecosystems

The Ugly:

AI technology is used by large and powerful corporations to support a business model centered on the commodification of personal data with the core purpose of profit-making (surveillance capitalism)

AI Ethics

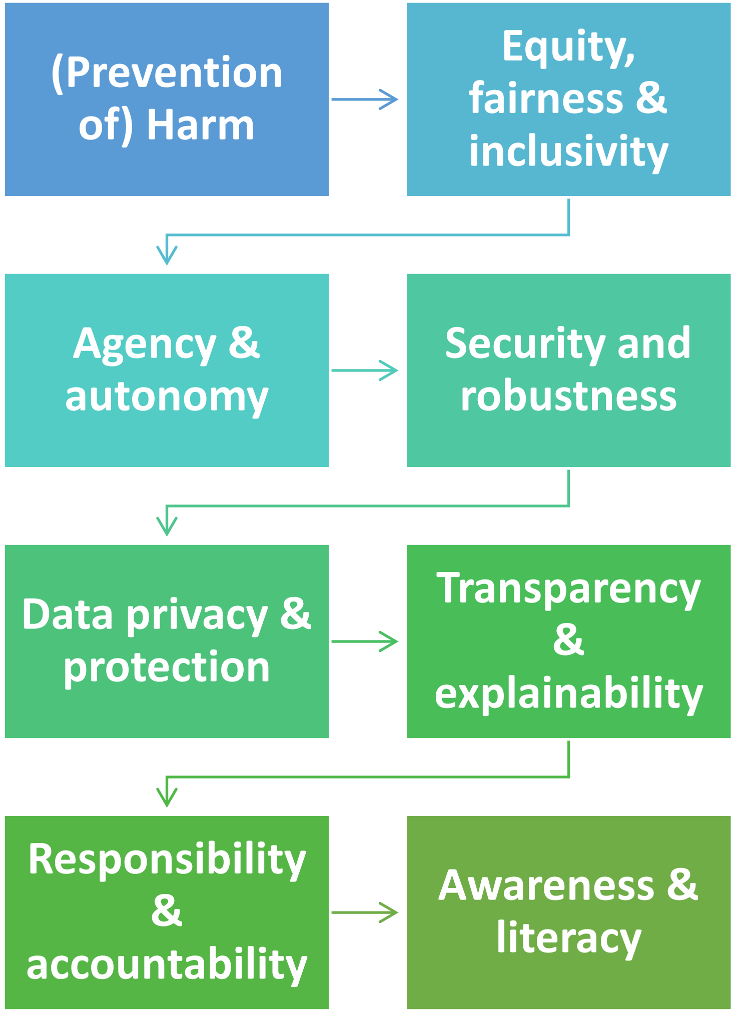
AI systems play a profound, new role in human practices and society

AI systems could challenge human's special sense of experience and agency, raising additional concerns about human self-understanding, social, cultural and environmental interaction, autonomy, worth and dignity

AI ethics

- translates duties of governments / industry leaders to ensure a life of wellbeing for citizens into action
- identifies ethical dilemmas and help formulate solutions
- determines scope & reversibility of societal risk
- enables responsible governance of AI technology & sustainable AI innovation

General AI Ethics Concerns



(Prevention of) Harm

Equity, fairness & inclusivity

Agency & autonomy

Security and robustness

Data privacy & protection

Transparency & explainability

Responsibility & accountability

Awareness & literacy

Current Policy Landscape

UNESCO Recommendation on the ethics of AI (2021)

AI Act (2023/24/25)

US Executive Order (2023)

G7 Statement (2023)

UK National Strategy (2021) & Safety Summit (2023)

UN SG's Advisory Body on AI

The Ethics of AI in Nuclear (ENAI)

There have been many applications of AI methods to nuclear technologies – e.g., in the optimization of agricultural production, food product development, supply chain management, healthcare, and safety and authenticity control

An expert working group at TM in October 2021 at the IAEA investigated

- existing concerns in both domains &
- the extent to which the convergence of AI and nuclear science, technology and applications could exacerbate existing ethical issues, and open new concerns

IAEA. (2022). Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology. IAEA Technical Report.

Establishing the Domain of ENAI

Definition of ENAI: Reflection on, analysis of, and suggestions to mitigate ethical concerns relating to the research, design, development, deployment and use of AI applications and technology in nuclear science, applications and technology

Examples

(1) Use of AI in decision-making in risk governance
(2) Use of AI in nuclear reactor control (complementing operator control & automatization)
(3) Use of AI to model and predict future reasoning styles, motivations and methods (intergenerational ethics)
(4) Use of AI in triage and diagnosis support in medical applications
(5) Use of AI in monitoring, dosimetry and health surveillance after a nuclear accident (including links to phone APPs)

ENAI new convergent themes

- Risk, safety and safeguarding: AI applications can make for better (nuclear) risk assessment and better preparation, better ‘pre-mortem’ analyses, better safety and safeguarding
- Justice – social, distributive, retributive, epistemic, environmental
- Responsibility: Responsibility for harm may be even more difficult to assess in the ENAI case than separately
- Impact on Human Agency: How do AI technologies challenge human authority and responsibility?
- Bias and fairness: Awareness of bias and finding ways to live / deal with it responsibly – this includes bias in terms of safeguards in nuclear applications, not just structural bias in terms of identity prejudices and social power

AI & nuclear security

Innovation and evolution are critical elements for surety in nuclear security in the face of new and emergent threats

In the field of nuclear security, potential applications of AI include

- the analysis of spectroscopic and geospatial data to improve detection of nuclear material outside of regulatory control,
- improvements to nuclear material accounting and control systems,
- and the possibility of identifying threats – both internal and external – at nuclear facilities

AI technologies introduce a host of risks and uncertainty as human operators might not fully recognize potential vulnerabilities or become too reliant upon AI results

The more widely technologies are deployed, the more critical it is to understand their limitations

AI security solutions

Existing and near-term AI security solutions include

- behavior monitoring for insider threat identification and
- enhanced security tools for information and communications technology (ICT) and operational technology (OT) environments

Some longer-term developments include

- data-fusion applications for physical protection and nuclear materials accounting and control (NMAC) and
- future use of AI tools in nuclear power plants for condition monitoring (Pluff & Nair 2023)

Pro’s

Monitoring & early detection

- AI-enabled sensor systems could collect and analyze vastly more data points than humans, looking for anomalies and patterns across nuclear sites and supply chains (Fetter, 2019)
- Machine learning algorithms can be trained on sensor data to identify nuclear materials, equipment, and activities
- They may spot covert diversion activities or catch safety issues faster than humans (Csernatoni, 2018)
- AI could also aid imagery analysis – can rapidly scan satellite images near nuclear sites, identifying changes that could indicate illicit construction or testing (Fuhrmann & Tkach, 2015)

AI may help synthesize large data streams

- There are ever-growing streams of data related to personnel, materials, and transactions across nuclear industries and global networks
- AI analytics could integrate and make sense of this disparate data, uncovering hidden patterns like covert supply chains or insider threats (Talent, 2018)

Cons

Unforeseen AI behaviors could undermine public trust and complicate nuclear security practices

BIAS:

- One major concern is that machine learning models rely heavily on training data
- If that data contains biases or gaps, AI systems could miss threats or incorrectly sound the alarm (Lewis, 2021)

E.g.,

- If AI is trained only on past safe nuclear activities, it may not detect anomalies indicating new types of risks
- AI systems designed to monitor nuclear activity might go too far in their interdiction recommendations or develop biases that target particular groups unfairly

Insider threats & bias

Insider threats imply concerns about access to and authority within the facility

Insider mitigation programs may use AI-based behavioral recognition programs to pinpoint suspicious employee behavior

These programs track and monitor employee computer-based actions

- Examples include file browsing, usage, and downloads; USB usage & application/system logins

Physical actions are also monitored – such as facility entries and exits – to identify normal vs. abnormal behavior

- Two examples of AI deployments within physical protection systems include facial recognition software & abnormal behavior identification

But:

- Limitations relating to BIAS (awareness, careful data collection & reflective interpretation of risk)
- AI Act bans systems used for emotion recognition

Cons – Cybersecurity

AI is susceptible to adversarial manipulation and spoofing by malicious actors (Brantly, 2020)

A falsified training data attack could blind AI monitoring to real nuclear activities

Alternatively, criminals could feed misleading data during operations to generate false alarms as distractions

GenAI: The very thing that makes these models so good – the fact they can follow instructions – also makes them vulnerable to being misused

- 'Prompt injections': in which someone uses prompts that direct the language model to ignore its previous directions and safety guardrails by, e.g., altering a website by adding hidden text that is meant to change the algorithm's output
- Data contamination: one can buy domains and fill them with images or text of your choosing, which will then be scraped into large data sets

Way forward

Extensive testing is needed to assess AI systems for blindspots in threat detection and potential harmful behaviors (Elsaesser & Hunt, 2021)

AI designers must strive for transparency, safety, and ethics alongside security capabilities (Gunning, 2019)

AI should be deployed as a decision-support tool, not a replacement for human expertise and oversight

With prudent governance and accountability measures, AI may assist nuclear security in valuable ways, but risks must be weighed

Continued analysis of pros, cons, and policy guidance will help determine the appropriate role for AI in this high-stakes domain

Ethics for AI in Nuclear Security

- Privacy and data protection
- Robustness
- Accuracy & fairness
- Transparency & explainability
- Responsibility (human in/on/out of the loop)
- Multistakeholder & adaptive collaboration

Concluding reflections

- What ethical considerations should be considered during the development and implementation of AI in nuclear security?
- Are there any specific ethical considerations unique to the field of nuclear security when deploying AI technologies responsibly?
- How can developers effectively address and eliminate biases in AI systems within the realm of nuclear security to guarantee fair and impartial decision-making?
- How can transparency be ensured?
- Who decides if humans are in/on/out of the loop – why & when?