Centre for Emerging Technology and Security

RESEARCH REPORT

OXFORD INTERNET INSTITUTE

Artificial Intelligence for National Security:

The Predictability Problem

Mariarosaria Taddeo, Marta Ziosi, Andreas Tsamados, Luca Gilli and Shalini Kurapati

September 2022

About CETaS and the Oxford Internet Institute

The Centre for Emerging Technology and Security (CETaS) is a policy research centre based at The Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to inform UK security policy through evidence-based, interdisciplinary research on emerging technology issues. Connect with CETaS at cetas.turing.ac.uk.

The Oxford Internet Institute – founded in 2001 – is a multidisciplinary research and teaching department of the University of Oxford, dedicated to the social science of the Internet. The Institute aims to shape the development of our digital world for the public good, operating at the cutting edge in both quantitative and qualitative methodologies that cut across disciplines and topics.

All views expressed in this report are those of the authors, and do not necessarily represent the views of The Alan Turing Institute, the Oxford Internet Institute, or any other organisation.

Acknowledgements

The authors are grateful to The Alan Turing Institute and CETaS for having supported the research presented in this report with both funding and expertise. We are also grateful to the many colleagues with whom we discussed the content presented in this report, their feedback helped us to refine several crucial aspects. We also would like to thank the reviewers, whose comments enabled us to sharpen our analysis and ensure clarity. We remain the only ones responsible for any remaining mistakes.

Executive Summary

This report focuses on the risks related to the potential lack of predictability of AI systems – referred to as the *predictability problem* – and its implications for the governance of AI systems in the national security domain. Predictability of AI systems indicates the degree to which one can answer the question: *what will an AI system do?* The predictability problem can refer both to correct and incorrect outcomes of an AI system, as the issue is not whether the outcomes follow logically from the working of the system, but whether it is possible to foresee them at the time of deployment.

There is growing concern that the use of unpredictable AI systems to inform high-stakes decisions may lead to disastrous consequences, which would undermine public trust in organisations deploying these systems and potentially erode the reputations of governments. In the national security domain, the use of AI introduces a new source of uncertainty that can hinder risk management procedures and potentially muddy the chain of accountability. In this domain, the implications of the predictability problem could lead to security risks for critical infrastructure, risks to the rights and well-being of individuals, conflict escalation or diplomatic fallout.

In this report, we first analyse the predictability problem from technical and socio-technical perspectives and then focus on relevant UK, EU and US policy to consider whether and how they address this problem. From a technical perspective, we argue that given the multi-faceted process of design, development, and deployment of an AI system, it is not possible to account for all sources of errors or emerging behaviours that could result. Moreover, even in an ideal scenario where no errors at design or development stage can be assumed or detected, once deployed an AI system may still develop formally correct (but unwanted) outcomes, which were not foreseeable at the time of deployment.

We analyse the socio-technical implications of the predictability problem by focusing on human-machine teams (HMT-AI). These teams represent an increasingly common mode of

deployment of AI systems. In HMT-AI, humans consult, coordinate with, rely on, develop and exchange tasks with AI agents. As HMT-AI combine human and artificial autonomy, they exacerbate the predictability problem by multiplying the number and types of interactions between artificial and human agents and their environment. We identify three main sources of the predictability problem in this context: human-machine interfaces, training of personnel, and (over)trust. Human-machine interfaces may foster unpredicted outcomes, insofar as they can conceal, distort or detail excessively the workings of AI systems, and training programs may not account for the learning capabilities of AI technologies and long-term convention building in HMT-AI. In the same way, over-trust dynamics whereby human agents in an HMT-AI accept uncritically the outcomes of AI systems may also lead to unpredicted results.

Having identified some of the root causes of the predictability problem, we analyse UK, EU and US policies, to assess whether these causes are covered in relevant policy documents and, if so, how and to what extent. We identified four main themes and a gap. These are: control, oversight, and value alignment; the resource boosting approach; the development of trustworthy AI; and the lack of focus on risk management measures to curtail the impact of the predictability problem.

Our policy analysis includes eight recommendations to mitigate the risks related to the predictability problem. The key suggestions are to centre governance approaches on HMT-AI rather than only AI systems and to conceptualise the predictability problem as multi-dimensional, with solutions focussed on shared standards and criteria for the composition of HMT-AI. Among these standards and criteria, requirements of trustworthy AI are particularly relevant and should be coupled with standards and certification schemes assessing the predictability of AI systems and procedures to audit HMT-AI. Cost-benefit analyses and impact assessments underpinning the decision to use HMT-AI in national security should account for the predictability problem and its potential impact on human rights, democratic values, and risk of unintended consequences. To ensure sufficient risk management when

deploying potentially unpredictable AI systems, we suggest adapting the ALARP principle – as low as reasonably practical – as a foundation for developing an AI-specific risk assessment framework of the predictability problem in HMT-AI.

The proposed ALARP-based framework would offer useful practical guidance, but alone would not be sufficient to identify and mitigate the risks posed by the predictability problem. Additional policy, guidance and training is required to fully account for the risks presented by the AI predictability problem. The higher the impact of the decisions that an AI system supports, the greater is the duty of care on those designing, developing, and using that system, and the lower the acceptable risk threshold. The analysis and recommendations should be read as actionable insights and practical suggestions to support relevant stakeholders to foster socially acceptable and ethically sound uses of AI in the national security context.

Recommendations

Recommendation 1. Government research funding should be allocated to develop public-private collaborations and longitudinal studies on HMT-AI. This research should focus on old and new models for decision-making in HMT-AI to assess the impact of team conventions building and training on performance and control measures. Focus should be drawn on defining new training protocols for HMT-AI specific dynamics, and on accelerating the development of risk management standards and HMT-AI performance assessments.

Recommendation 2. A dedicated certification scheme for HMT-AI should be established, to promote industry consensus on the design requirements and evaluation of AI systems designed for HMT-AI. Generalising between tasks, effective communication, performance consistency, and adapting to new teammates should all be included within such a certification scheme. Building on under-developed ISO standards, this certification scheme should also extend to the traceability of processes and decision accountability as well as auditing mechanisms to evaluate levels of trust in HMT-AI. This is necessary to disincentivise

over-trust and complacent attitudes in HMT-AI that maintain or amplify the predictability problem.

Recommendation 3. Policy responses to the predictability problem in the national security domain should focus on governing HMT-AI teams, rather than AI systems alone.

Recommendation 4. Cost-benefit analyses (CBA) of HMT-AI in the national security domain should include an assessment of the predictability of AI systems and of the related ethical risks along the technical and operational dimensions. To facilitate coherent assessment across security agencies, a standard scale to assess predictability of AI systems should be defined, where the choice of using (or not) AI should be justified on this scale with respect to a contextual CBA as well as the consideration of public attitudes towards the risks and the benefits involved. The definition of this scale should be within the remit of an independent third-party actor, i.e., a different public office than the one deploying the HMT-AI.

Recommendation 5. Rather than "more" or "less" predictability, policy proposals should focus on predictability trade-offs, making clear which aspect of the predictability problem specific proposals aim to tackle and in which way, as well as which aspects they risk exacerbating, and which mitigating measures will be put in place. Policies should recognise that predictability is a multi-dimensional concept, where gains in predictability on one level can come at the expense of losses on another.

Recommendation 6. Policies on the problem of AI predictability in national security should address the link between trustworthiness and unpredictability, both at a formal and operational level. For example, AI systems should be given an amendable predictability score, which should be included in the assessment of the trustworthiness of the system. The trustworthiness of an AI system should include a cost-benefit analysis to assess the risks that unwanted behaviour may pose in different contexts of deployment.

Recommendation 7. Risk thresholds should be established for unpredictable AI which map the severity of risks around unpredictable behaviour to their own level of predictability (e.g., division into known knowns, known unknowns, etc.). These thresholds will in turn inform the development of risk management processes, allowing risks to be prioritised based on their predictability and their impact.

Recommendation 8. An ALARP-based framework should be developed to assess the risks of unpredictable AI and HMT-AI, and establish the maximum acceptable degree of unpredictability for any given context. This framework should include:

- A quantitative assessment of the level of predictability of a given AI system and HMT-AI;
- An assessment of the traceability of the design, development, and/or procurement steps leading to deployment of the AI system;
- An assessment of the conditions of deployment, e.g., HMT-AI, level of training of operators (or HMT-AI members), level of transparency of the interface, level of human control over the AI system;
- A cost-benefit analysis of the potential risks and intended benefits of deploying the system (as per Recommendation 4);
- An analysis of hypothetical scenarios to consider how exposure to risk or the effectiveness of mitigating measures may vary with context of deployment;
- Protocols for human overriding of the system and redress mechanisms.

1. Introduction

Artificial Intelligence (AI) is becoming a key element of contemporary security organisations.[1] Recent advances in AI, such as deep learning (DL), have triggered a wave of research and development in the field, and an uptake in experimentation with AI systems in various security settings.[2] AI is now considered a key technology for maintaining advantage over adversaries and protecting against threats.[3] The UK Government Communications Headquarters (GCHQ) has recently stated that 'AI capabilities will be at the heart of our future ability to protect the UK'.[4] In the USA, the National Security Commission on Artificial Intelligence stated that 'AI will revolutionize the practice of intelligence', and that 'there may be no national security function better suited for AI adoption than intelligence tradecraft and analysis'.[5]

There are several potential uses of AI across different national security contexts, including but not limited to: the use of AI for the automation of administrative and organisational processes; the use of AI for cybersecurity processes; and the use of AI for intelligence analysis – otherwise known as 'augmented intelligence', which could include automated analysis of text, image or audio data, filtering of content derived from bulk collection, or behavioural analytics at the individual person level. These uses pose similar ethical challenges to those emerging in other domains (Figure 1). Problems concerning attribution of responsibility, lack of transparency, fairness and bias, assessment of justified, proportionate and necessary uses, and control over AI systems, for example, have been

[1] Lewis, Larry. 'Resolving the Battle over Artificial Intelligence in War'. *The RUSI Journal* 164, no. 5–6 . pp.62-71. 19 September 2019.; Stevens, Tim. 'Knowledge in the Grey Zone: AI and Cybersecurity'. *Digital War* 1, no. 1. pp.164-170. 1 December 2020.

[2] Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, and Christian Curriden. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. RAND Corporation, 2020.

[3] Taddeo, Mariarosaria, 'Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity'. *Minds and Machines* 29, no. 2. pp.187-191. June 2019.

[4] GCHQ, 'Pioneering a New National Security: The Ethics of Artificial Intelligence'. p.4. 2021.

[5] NSCAI. 'Final Report'. Washington DC: National Security Commission on Artificial Intelligence. p.23. 2021.

reported and analysed in defence,[6] healthcare,[7] finance,[8] and all remain relevant when considering the use of AI for national security purposes.

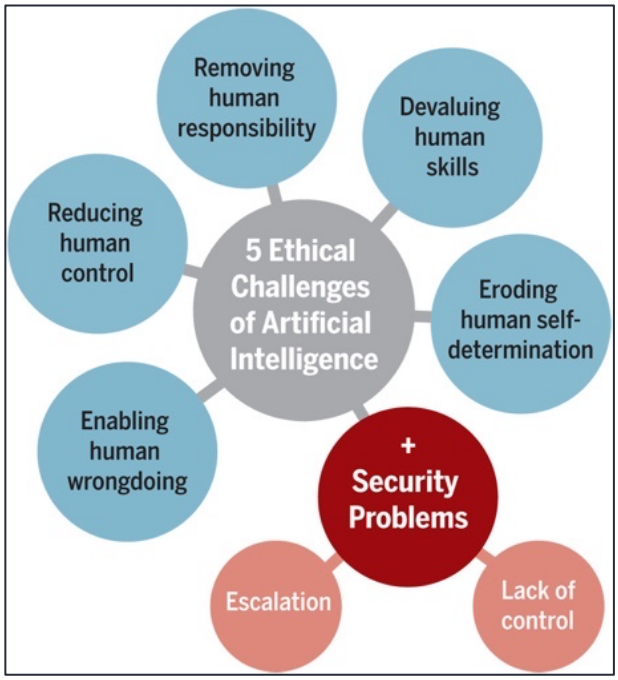

Figure 1. The ethical and security challenges coupled with the use of AI systems, figure from Yang et al. (2018).[9]

In this report, we focus on the risks of the potential lack of predictability of the outcomes of AI systems – referred to as the *predictability problem* - and its implications for the governance of AI systems in the national security domain. The predictability problem is conceptually preeminent to the challenges described above. Unpredictable AI systems pose challenges for anticipating their effects whether intended or not, for ensuring control, protecting human autonomy and judgement when interacting with AI systems, and for ascribing responsibility and accountability for the decisions made on the basis of AI outputs.[10] When unpredictable

[6] Taddeo, Mariarosaria, David McNeish, Alexander Blanchard, and Elizabeth Edgar. 'Ethical Principles for Artificial Intelligence in National Defence'. *Philosophy & Technology* 34, no. 4. pp.1707-1729. 1 December 2021.
[7] Morley, Jessica, Caio C.V. Machado, Christopher Burr, Josh Cowls, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi. 'The Ethics of AI in Health Care: A Mapping Review'. *Social Science & Medicine* 260: 113172. September 2020.
[8] Svetlova, Ekaterina. 'AI Ethics and Systemic Risks in Finance'. *AI and Ethics*. 13 January 2022.
[9] Yang, Guang-Zhong, Jim Bellingham, Pierre E. Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, et al. 'The Grand Challenges of Science Robotics'. *Science Robotics* 3, no. 14. p.10. 31 January 2018.
[10] Taddeo, Mariarosaria, and Alexander Blanchard. 'Ascribing Moral Responsibility for The Actions of Autonomous Weapons Systems: A Moral Gambit'. *SSRN Electronic Journal*. 2022.

AI systems are used to inform high-stakes decisions, such as those concerning national security, uncertainties inherent in the systems may jeopardise individuals' and groups' fundamental rights.[11] In turn, this could undermine public trust in organisations deploying these systems and, when organisations belong to the public sector, erode the reputation of governments.

In the national security domain, the use of AI systems in intelligence operations,[12] counterterrorism,[13] law enforcement,[14] computer network operations,[15] and military activities[16] introduces levels of uncertainty that may hinder risk management procedures or muddy chains of decision-making accountability.[17] In this domain, the implications of the AI predictability problem may lead to security risks for critical infrastructure, risks to the rights and well-being of individuals, and could also lead to conflict escalation and diplomatic fallout.

Crucially, the problem of predictability entails a necessary and unavoidable degree of uncertainty with respect to possible outcomes of an AI system. Reducing this uncertainty is key when considering the use of AI systems to inform high-impact decisions. The higher the impact of the decisions that an AI system supports, the greater the duty of care of those designing, developing, and using that system, and the lower the acceptable risk threshold.

---

[11] Tsamados, Andreas, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. 'The Ethics of Algorithms: Key Problems and Solutions'. *AI & SOCIETY*. 20 February 2021.

[12] Baber, Chris, Ian Apperly, and Emily McCormick. 'Understanding The Problem Of Explanation When Using AI In Intelligence Analysis', 2021.

[13] UNCCT, United Nations Office of Counter Terrorism. 'Countering Terrorism Online With Artificial Intelligence', 2021.

[14] Babuta and Oswald, 'Data Analytics and Algorithms in Policing in England and Wales'. p.62. 2020.

[15] Stevens, Tim. 'Knowledge in the Grey Zone: AI and Cybersecurity'. *Digital War* 1, no. 1. pp.164-170. 1 December 2020.

[16] Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, and Christian Curriden. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. RAND Corporation. 2020.

[17] NATO, ed. *NATO Code of Best Practice for Command and Control Assessment = Code OTAN Des Meilleures Pratiques Pour l'??Valuation Du Commandement et Du Contr??Le.* Neuilly-sur-Seine Cedex, France: North Atlantic Treaty Organisation, Research and Technology Organisation, 2004; Parra-Arnau, Javier, and Claude Castelluccia. 'Dataveillance and the False-Positive Paradox'. April 2018.

The analysis and recommendations offered in the following sections should be read as actionable insights and practical suggestions to support relevant stakeholders to this end.

This report is structured as follows. Section 2 defines the predictability problem. Sections 3 and 4 analyse some of its root causes by exploring both technical and socio-technical aspects of the predictability problem. The reader not interested in these aspects may go directly to Section 5, where we analyse UK, EU, and US policies focusing on the use of AI for national security and assess whether and how they address the predictability problem, and offer recommendations to fill relevant gaps in the existing governance approaches. Section 6 concludes our analysis. The Appendix provides a glossary defining key terms used in this report.

2. The Predictability Problem

Predictability of AI systems indicates the degree to which one can answer the question: *what will an AI system do?* Unpredictable systems are not a new issue. They are common in mathematics and physics, and limits on the ability to predict the outcomes of artificial systems have been proven formally since the 1950s.[18] Wiener and Samuel debated over the predictability of AI systems in a famous exchange in 1960.[19] Wiener attributed the lack of predictability to the learning abilities of these systems, noting, "as machines learn they may develop unforeseen strategies at rates that baffle their programmer".[20]

Developments in AI research have proved Wiener correct. Consider, for example, reward hacking, which is reported in current literature as one of the factors that can make an AI system unpredictable:

> "Autonomous agents optimize the reward function we give them. [...] When designing the reward, we might think of some specific training scenarios, and make sure that the reward will lead to the right behavior in *those* scenarios. Inevitably, agents encounter *new* scenarios (e.g., new types of terrain) where optimizing that same reward may lead to undesired behavior".[21]

[18] Rice, H. G. 'On Completely Recursively Enumerable Classes and Their Key Arrays'. *Journal of Symbolic Logic* 21, no. 3. pp.304-308. September 1956; Musiolik, Thomas Heinrich, and Adrian David Cheok, eds. *Analyzing Future Applications of AI, Sensors, and Robotics in Society:* Advances in Computational Intelligence and Robotics. IGI Global. 2021.

[19] Wiener, N. 'Some Moral and Technical Consequences of Automation'. *Science* 131, no. 3410. pp.1355-1358. 6 May 1960.

[20] Wiener, N. 'Some Moral and Technical Consequences of Automation'. *Science* 131, no. 3410. p.1355. 6 May 1960.

[21] Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. 'Inverse Reward Design'. *ArXiv:1711.02827 [Cs]*. 7 October 2020.

Currently, predictability of AI systems is debated both at a technical and an operational level. Some AI researchers focus on the technical features of a system,[22] while others consider predictability a function of the system and its context of deployment, i.e., operational predictability.[23]

From a technical standpoint, predictability of an AI system is assessed in terms of the degree of consistency between its past, current, and future behaviours.[24] Key aspects monitored here are data and concept shift; how often and for how long the outputs of a system are correct; and whether the system can scale up to elaborate data that diverge from training and test data.[25] [26] Predictability also depends on properties such as interpretability, transparency, explainability and trustworthiness[27] of an AI system (discussed further in Section 3).

Predictability also refers to the degree to which the actions of a system can be anticipated once it is deployed in a specific context. In this sense, 'all autonomous systems exhibit a degree of inherent operational unpredictability, even if they do not fail or the outcomes of their individual action can be reasonably anticipated'.[28] Operational predictability is impacted by a large set of variables: the technical features of the system (e.g. whether it is an online or offline

[22] International Committee of the Red Cross, ICR. 'Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control'. 2019; Boulanin et al., 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control'; DIB, 'AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense - Supporting Document'.

[23] International Committee of the Red Cross, 'Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control'; Docherty, Bonnie. 'The Need for and Elements of a New Treaty on Fully Autonomous Weapons'. *Human Rights Watch*. 1 June 2020.

[24] Holland Michel, Arthur. 'The Black Box, Unlocked | UNIDIR'. 2020.

[25] Boulanin et al., 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control'; Collopy, Paul, Valerie Sitterle, and Jennifer Petrillo. 'Validation Testing of Autonomous Learning Systems'. *INSIGHT* 23, no. 1. pp.48-51. March 2020; DIB, 'AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense - Supporting Document'.

[26] It is important to note that predictability is not reliability (the degree of failures of a system) nor is it robustness (the capacity of a system to behave as expected even when it is fed with erroneous data); Heaven, Douglas. 'Why Deep-Learning AIs Are so Easy to Fool'. *Nature* 574, no. 7777. pp.163-166. 10 October 2019.

[27] Holland Michel, 'The Black Box, Unlocked | UNIDIR'. 2020; Rudin, Cynthia, Caroline Wang, and Beau Coker. 'The Age of Secrecy and Unfairness in Recidivism Prediction'. *Harvard Data Science Review* 2, no. 1. 31 March 2020.

[28] Holland Michel and Holland Michel, 'The Black Box, Unlocked'5. 2020.

learning system), the characteristics of the context of deployment, interactions with other systems, the level to which the operator understands the way in which the system works and, in the security domain, the behaviour of adversaries. These variables may change and interact in different ways making it problematic to predict all possible actions that an AI system may perform and their resulting effects.

In this report we define the predictability problem as follows:

> *Maximally*, given the multi-faced processes of design, development, and deployment of AI systems, the opaqueness of these systems, their adapting capabilities, and the possible complexities of the environment of deployment, it is neither possible to account for all sources of errors and manipulation of a system nor for all possible emerging behaviours – whether beneficial or not – of an AI system that these errors may prompt.
>
> *Minimally*, given an ideal scenario where no errors at design and development stages can be assumed or detected, once deployed an AI system may still develop correct (and yet unwanted) outcomes, which were not foreseeable at the time of deployment.

This definition allows us to stress that the predictability problem refers both to correct and incorrect outcomes, as in both cases the issue is not whether the outcomes follow logically from the working of an AI system, but whether it is possible to foresee them at the time of deployment.

It is also important to note that the unpredictability of an AI system is not boundless, rather it is limited by the system affordances – the set of hardware and software specifications that determine the range of possible actions of a machine. For example, an unsupervised system designed and developed to distinguish pictures of horses from those of dogs will be unpredictable with respect to the elaboration of visual inputs it will consider, the execution strategy, and the final selection of pictures. There is no concern that the system will develop an unpredicted behaviour outside its affordances and produce a new type of outcome, like

drawing a picture of a horse or a dog. It follows that given an operational context, the more complex the affordances of a system, the wider the range of unpredictable behaviours that it may show upon deployment. In the following sections, we describe some of the root causes of the predictability problem looking first at its maximal and then at the minimal definition. The reader interested in the governance implications of the predictability problem for the use of AI in national security may move directly to Section 5.

3. Root Causes of the Maximal Predictability Problem

The predictability problem impacts different types of AI in different ways depending on the type of learning model considered. For example, AI systems based on offline models are more predictable than AI systems based on online learning models. This is because offline models are trained with data in batches, while online models are continuously re-trained with live data. In this section, we offer a high-level description of the key aspects of AI systems that may lead to unpredictable outcomes. We refer to them as to the *root causes* of the predictability problem as described in the maximal definition in Section 2 and summarise them in Section 3. Our aim is not to provide an exhaustive list of root causes but to bring the reader's attention to a range of factors that affect the predictability of AI systems.

An AI system is built from different technology blocks, the set of which is referred to as the AI stack. There is no 'one-size-fits-all' technical stack. Different use cases and contexts have different requirements and determine different stacks. Table 1 below offers an example of an AI stack and its main building blocks. For each block, it also provides a list of the main root causes of the predictability problem. The following subsections focus on some of these causes related to the type of machine learning (ML) used, the data, and the practices underpinning the design and development of an AI system.

AI Stack Component	Description	Examples of root causes of the predictability problem
Computational Power	Virtual machines, physical servers, serverless options, and specialised hardware and container options. These may be self-hosted on-premises or cloud-based.	All computational platforms are vulnerable to hacking to some degree. Computer hardware, firmware, operating systems, and cloud infrastructure may all be vulnerable to bugs which may cause incorrect results. They are also reliant on stable power and internet connectivity and can be taken offline if one of these fails.
Input Data	Crucial input to ML systems. Data quality issues such as correctness, timeliness, and adequate coverage of the problem domain impact the outcomes and suitability of the ML model.	Data may be affected by data scarcity, label ambiguity and the inability to represent real-world scenarios and social bias replicated by humans. Processes such as data cleaning are resource intensive and laborious. There are no efficient data quality control and governance mechanisms for big data.
Machine Learning Platforms	Platforms necessary for developing machine learning capabilities. Many ML frameworks and libraries are available, supporting different ML algorithms and programming languages (Subramanian 2018). Cloud ML services are also available, such as Amazon Rekognition, SageMaker and Google Cloud's Vertex AI.	Machine learning platforms allow users to mix and match ready-made tools, models, datasets, or libraries that developers would then depend on, regardless of their ability to understand, modify or fix them when specific issues emerge.

Machine Learning Algorithms / Types of Learning	Supervised learning - Algorithms predict outputs based on a training dataset of labelled examples (usually labelled by humans).	From dataset-related biases to specification gaming to brittleness and adversarial examples. Explored more in the following section.
	Unsupervised learning - Algorithms do not rely on human labelling of training data. They are mainly used for exploring datasets, finding anomalies, patterns or clusters, and for determining features to be used in supervised learning applications.	From dataset-related biases (beyond labelling) to specification gaming to brittleness and adversarial examples. Explored more in the following section.
	Semi-supervised learning - Algorithms rely on a dataset in which only a subset of the training data is labelled. This type of ML relies heavily on the human labelling and the approval of the unknown data labelling.	From dataset-related biases to specification gaming to brittleness and adversarial examples. Explored more in the following section.
	Reinforcement learning - Algorithms are concerned with maximising some objective function subject to constraints.	Reward hacking or specification gaming, in which the agent produces an unexpected behaviour to maximise the objective function. The agent generates a solution that strictly abides by the stated objective in a way unintended by the human developer.

User Interface	Interactions between the system and the end user/operators. This may range from simple displaying of output to a more sophisticated process that allows expert users to modify the system's configuration in response to its performance.	Depending on the features of a given interface, it can foster different levels of mismatch between system behaviour and user practice. Discussed further in Section 4.1.
Monitoring Solution	Essential for ensuring that the data used to train the model appropriately represents the live data. The monitoring module should quantify data and concept drift, prevent repeated prediction errors, and define re-training strategies.	Alignment problems. Failing to detect data and concept drift might make the model operate in conditions outside its design space, causing unpredictability.

Table 1. Main technology components of an AI Stack and examples of root causes of the predictability problem.

3.1 Machine Learning

ML can be defined mathematically, statistically or algorithmically.[29] Unifying these, ML corresponds to building complex functions to create a mechanism for pattern searching, by which a system can learn to identify patterns when presented with unseen data or scenarios. One of the main issues associated with the best performing families of ML models (like networks and boosted trees) is that their complexity makes it increasingly difficult to assess whether models are generalising appropriately on data outside training distributions. Model confidence is the most common approach in modern ML to deal with uncertainty associated, among other things, with generalisation. It assesses the different uncertainties characterising the model and its operational environment.[30] However, model confidence is often not

[29] Gollapudi, Sunila. *Practical Machine Learning*. Packt Publishing Ltd. 2016

[30] Hüllermeier, Eyke, and Willem Waegeman. 'Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods'. *Machine Learning* 110, no. 3. pp.457-506. March 2021.

statistically robust. Deep neural networks, for example, have been proven to be overconfident, possibly leading to high-confidence mistakes and/or accidently concealing adversarial attacks being conducted on the model.[31] Confidence levels will have to be adjusted to outputs, and this complicates (and may perturbate) subsequent processes.

At the same time, even for the best performing AI models, training outcomes are not necessarily indicative of the capabilities of a system in the real world, where deployment conditions will diverge from training conditions, and new data falls outside the training dataset's distributions. Examples showing deep neural network models failing to generalise appropriately outside training conditions are extensively reported in the literature. For example, in computer vision – a popular application of AI – it is difficult to analyse images where there is a noisy context or contextual confusion of extraneous pixels or light. AI systems have been shown to be susceptible to minor changes, down to pixel level, with minor variations leading a system to misidentify 3D printed turtles as rifles [32] or stripes as school buses.[33] These limitations have been shown to be exploitable in high-stakes situations, like industrial settings where AI-enabled robotic arms have been tricked into harming human operators or in AI-enabled multi-domain defence operations.[34]

[31] ENISA, 'Artificial Intelligence Cybersecurity Challenges'. Report/Study. 2020.

[32] Athalye, Anish, Nicholas Carlini, and David Wagner. 'Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples'. *ArXiv:1802.00420 [Cs]*. 1 February 2018.

[33] Nguyen, Anh M., J. Yosinski, and J. Clune. 'Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images'. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[34] Jia, Yifan, Christopher M Poskitt, Jun Sun, and Sudipta Chattopadhyay. 'Physical Adversarial Attack on a Robotic Arm'. p.8. 2022; Savas, Onur, Lei Ding, Teresa Papaleo, and Ian McCulloh. 'Adversarial Attacks and Countermeasures against ML Models in Army Multi-Domain Operations'. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, 1141. SPIE. pp.235-240. 2020.

3.2 Data

Data is fundamental for training and deploying AI models. If data is fuel for AI, a data pipeline consists of the various processes ranging from data acquisition, transformations and storage before feeding into AI models. In this pipeline, data preparation or curation is a crucial step. Labelling is a key aspect of AI data curation when considering the predictability problem. Labels or tags[35] attach meaning to the data, enabling a machine to learn from it. For example, raw data can be labelled according to their type (e.g., photo) and their content (e.g., photo of a horse). Different methodologies are available for labelling, with important limitations that may lead to unpredicted system outcomes. When labelling requires human intervention, this introduces risks. For example, in-house or outsourced data labellers may reproduce bias,[36] creating skewed training which will impact the performance of the AI system and lead to unpredicted outcomes. Other forms of labelling, like consensus voting labelling, may improve overall labelling quality but at higher costs than other forms of labelling.

In some cases, it may be possible to create synthetic labelled data. Synthetic data may be created either from scratch using simulation techniques or can be generated based on real world 'seed' data with generative AI models. This often requires vast processing power and comes with elevated error potential. Errors may be introduced through (human) expert opinions in case of simulations, while bias and imbalances from the real-world seed data are propagated through generative models Although synthetic data is modelled on real-world distributions, the sample used for generation may not be representative. The resultant synthetic data may inherit any underlying biases or skew present in the sample data, and any

[35] Godwin, Jamie, and Peter Matthews. 'Robust Statistical Methods for Rapid Data Labelling'. Chapter. Data Mining and Analysis in the Engineering Field. IGI Global. 2014.

[36] Bekele, Esube, Cody Narber, and Wallace Lawson. 'Multi-Attribute Residual Network (MAResNet) for Soft-Biometrics Recognition in Surveillance Scenarios'. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. Washington, DC, DC, USA: IEEE. pp.386-393. 2017'; Bekele, Esube, Wallace E Lawson, Zachary Horne, and Sangeet Khemlani, 'Human-Level Explanatory Biases for Person Re-Identification'. p.2. 2018.

downstream supervised learning process may then learn and replicate those unrepresentative distributions– leading to unpredictable outcomes.

Data cleaning is another form of curation. It removes duplicates, missing values, uninformative features, and outliers from a dataset under the assumption that they represent incorrect data or an error.[37] The goal is to improve model performance. However, data cleaning introduces a risk of removing meaningful data points which may be important in the application phase. Removing this data may lead to unforeseen and unintended outcomes since the resulting 'clean' data set may be stripped of important information, useful for testing model behaviour.

When datasets are constructed from multiple sources, some of which may be incompatible, the data commingling problem emerges. This is the case, for example, when different sensors may be used without having been calibrated or normalised to produce the same values. This may lead to inconsistent, incomplete, or inaccurate datasets, and in turn to unreliable outcomes.

Data shift (also referred to as data drift) is the extent to which system outcomes have moved off-course due to external factors leading to a change in data distribution.[38] Building an AI model requires identifying predictable relationships between input and target variables. The expectation is that the same data distribution would elicit similar results. However, real-world examples rarely show this to be the case,[39] where various unpredictable factors can change input, dataset quality, data capture (for example polling frequency), or even the underlying patterns forming relations between input and output data.

[37] Tobin, Donal. 'What Is Data Cleansing and Why Does It Matter?' Integrate.io. 2022.
[38] Sarantitis, George. 'Data Shift in Machine Learning: What Is It and How to Detect It'. *Georgios Sarantitis* (blog). 16 April 2020.
[39] Sarantitis, George. 'Data Shift in Machine Learning: What Is It and How to Detect It'. *Georgios Sarantitis* (blog). 16 April 2020.

In addition to introducing unwanted errors in the outputs of an AI system, data curation steps present two important operational challenges that may increase the likelihood of errors leading to unpredictable behaviour of the system. The first is the operational pay-off (efforts vs efficiency) of conducting data curation; the second emerges from the lack of standards and automated mechanisms to evaluate data quality. Key dimensions of data quality include completeness, accuracy, uniqueness, timeliness, consistency, and validity (see Glossary). However, these dimensions and their relative importance may vary depending on the context of use and the related purpose. While data quality standards and governance mechanisms are relatively uncontested for structured data, this is not the case for unstructured data,[40] which accounts for the majority of data used in AI models.[41]

While the amount of data available continues to grow, there is a lack of agreed standards, tools and mechanisms[42] to evaluate data robustness continuously, and check whether it is fit for purpose. Limits in assessing data quality could lead to noise, errors, and inconsistency in data sets, and these may lead to unpredicted behaviour at the system level. These data-related errors and uncertainties, if unchecked, can continue to accrue and propagate across the various elements of the AI stack, as shown in Table 1. This brings us to issues related to technical debt.

3.3 Technical Debt

In software development, 'technical debt' is a metaphor used to refer to long-term software issues and costs stemming from forgoing best practices at the development stage in favour of easier and quicker solutions. Best practices commonly implemented in modern software development – like version control and unit and system testing – are not so easily translated to the AI domain due to a lack of standard procedures and frameworks, and the inherent

[40] Further information: https://arxiv.org/abs/1803.09010s

[41] Further information: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf

[42] Further information: https://datacentricai.org/data-in-deployment/

difficulty of defining robust tests for AI models.[43] Mitchell et al.,[44] among others, have proposed solutions such as packaging and shipping production models using model cards, which describe quantitatively the model's design space, key metrics, and known limitations, but these have yet to see widespread adoption.

The lack of commonly accepted versioning and testing tools in the AI domain generally translates into a scarce adoption of Continuous Integration/Continuous Delivery (CI/CD) practices, commonly used in software development to ensure that frequent code changes do not interfere with other changes made by developers working in parallel. Reliable CI/CD pipelines require, for example, extensive versioning. Not being able to version an AI system reliably can cause robustness issues in deployment phases and is a cause of the predictability problem in its maximal definition. At the same time, the numerous and interrelated components of an AI system make its abstraction boundaries hard to control. These aspects become more pressing when ML models continue to evolve upon deployment. When it occurs, technical debt hinders the reliability and traceability of the behaviour of an AI system, and in turn limits the ability of an observer to predict its outcomes.

[43] Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 'Hidden Technical Debt in Machine Learning Systems'. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc. 2015.
[44] Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 'Model Cards for Model Reporting'. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. pp.220-29. 2019.

4. Root Causes of the Minimal Predictability Problem

This section explores the socio-technical implications of the predictability problem through a focus on Human Machine Teams (HMT), where machines are AI systems (HMT-AI). Here, our analysis centres on the minimal definition of the predictability problem (see section 2). That is, we assume that the AI systems in question have been designed and developed to be as performant and robust as possible, and that unwanted outcomes hereinafter can be explained in view of the operational and human realities of deployment. To do so, we embrace a socio-technical approach, that is we focus on both technical and non-technical factors, i.e., cultural, ethical, legal and cognitive, to map the causes of the predictability problem, minimally defined.[45] [46]

HMT-AI mark a pivot from previous approaches to AI deployment, which assumed a clear division of labour between human and artificial agents (machines, including AI agents), rested on low levels of automation, and ascribed the processing of multiple sources of information

[45] Ehsan, Upol, and Mark O. Riedl. 'Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach'. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, edited by Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones. pp. 449–66. Lecture Notes in Computer Science. Cham: Springer International Publishing. 2020.

[46] Andras, Peter, Lukas Esterle, Michael Guckert, The Anh Han, Peter R. Lewis, Kristina Milanovic, Terry Payne, et al. 'Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems'. *IEEE Technology and Society Magazine* 37, no. 4. pp.76-83. December 2018; Chopra, Amit K., and Munindar P. SIngh. 'Sociotechnical Systems and Ethics in the Large'. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. pp.48–53. AIES '18. New York, NY, USA: Association for Computing Machinery. 2018; Ehsan, Upol, and Mark O. Riedl. 'Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach'. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, edited by Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones. pp. 449–66. Lecture Notes in Computer Science. Cham: Springer International Publishing. 2020; Makarius, Erin E., Debmalya Mukherjee, Joseph D. Fox, and Alexa K. Fox. 'Rising with the Machines: A Sociotechnical Framework for Bringing Artificial Intelligence into the Organization'. *Journal of Business Research* 120. pp.262-273. November 2020; NIST, 'AI Risk Management Framework: Initial Draft'. 2022.

only to humans.[47] Today, research on HMT-AI focuses on producing such 'joint-intelligence' systems, whereby the tasks of human experts and AI systems are distributed to create flexible team processes and facilitate emergent capabilities.[48] HMT-AI characterise the deployment of AI systems in several domains[49] from warehouse facilities,[50] urban search-and-rescue teams and advanced surgical operations teams,[51] to cybersecurity[52] and defence operations.[53]

As HMT-AI combine human and artificial autonomy, they exacerbate the predictability problem by increasing the amount and types of interactions (and sources of perturbations) between artificial and human agents, and their environment.[54] In this section we analyse three

[47] Shaw, Tyler, Adam Emfield, Andre Garcia, Ewart de Visser, Chris Miller, Raja Parasuraman, and Lisa Fern. 'Evaluating the Benefits and Potential Costs of Automation Delegation for Supervisory Control of Multiple UAVs'. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 54, no. 19. pp.1498-1502. September 2010; Walliser, James C., Ewart J. de Visser, Eva Wiese, and Tyler H. Shaw. 'Team Structure and Team Building Improve Human–Machine Teaming With Autonomous Agents'. *Journal of Cognitive Engineering and Decision Making* 13, no. 4. pp.258-278. December 2019; Woods, D. D., E. S. Patterson, and E. M. Roth. 'Can We Ever Escape from Data Overload? A Cognitive Systems Diagnosis'. *Cognition, Technology & Work* 4, no. 1. pp.22-36. 1 April 2002.

[48] O'Neill, Thomas, Nathan McNeese, Amy Barron, and Beau Schelble. 'Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature'. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. 22 October 2020.

[49] Lavin, Alexander, Hector Zenil, Brooks Paige, David Krakauer, Justin Gottschlich, Tim Mattson, Anima Anandkumar, et al. 'Simulation Intelligence: Towards a New Generation of Scientific Methods'. *ArXiv:2112.03235 [Cs]*. 6 December 2021; Scherrer, Nino, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C. Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. 'Learning Neural Causal Models with Active Interventions'. *ArXiv:2109.02429 [Cs, Stat]*. 5 March 2022.

[50] Stowers, Kimberly, Lisa L. Brady, Christopher MacLellan, Ryan Wohleber, and Eduardo Salas. 'Improving Teamwork Competencies in Human-Machine Teams: Perspectives From Team Science'. *Frontiers in Psychology* 12. 24 May 2021: 590290.

[51] You, Sangseok, and Lionel Robert. 'Emotional Attachment, Performance, and Viability in Teams Collaborating with Embodied Physical Action (EPA) Robots'. 2016.

[52] Stevens, Tim. 'Knowledge in the Grey Zone: AI and Cybersecurity'. *Digital War* 1, no. 1. pp.164-170. 1 December 2020.

[53] Konaev, Margarita, and Husanjot Chahal. 'Building Trust in Human-Machine Teams'. 2021.

[54] Lavin, Alexander, Ciarán M. Gilligan-Lee, Alessya Visnjic, Siddha Ganju, Dava Newman, Atılım Güneş Baydin, Sujoy Ganguly, et al. 'Technology Readiness Levels for Machine Learning Systems'. *ArXiv:2101.03989 [Cs]*. 29 November 2021

aspects of HMT-AI that have an impact on the predictability problem: human-machine interfaces, training, and trust.

4.1 HMT-AI and Human-Machine Interface

Human machine interface is one of the main areas of research in the field of HMT. It is a crucial area of development for defence agencies, and has been listed as a key priority for the DoD.[55] It is an area of research particularly relevant in the national security domain. An example is the use of AI for augmented intelligence,[56] where effective human-machine interfaces for bulk data analysis and predictive analytics may allow human agents to search and understand high dimensional data that would otherwise remain untapped.[57]

The design of human-machine interfaces aims to foster interactive, bi-directional processes. It structures information to give human operators situational awareness and can enable real-time human contributions to the AI agent's inferences or post-operation calibration. Effective human-machine interfaces leverage expert feedback, labelling and other types of human input to improve the AI agent's performance in real-time or in batches. For example, human feedback can help a deep reinforcement learning system train for complex and novel behaviours, necessary to navigate real-world environments.[58] In cybersecurity, researchers combine experts' experience and intuition with machine learning techniques to create a system capable of detecting and defending against unseen attacks.[59]

---

[55] Lopez, Todd. 'Simplified Human/Machine Interfaces Top List of Critical DOD Technologies'. 2022.
[56] Babuta, Alexander, Marion Oswald, and Ardi Janjeva. 'Artificial Intelligence and UK National Security: Policy Considerations'. Occasional Paper. London: Royal United Services Institute for Defence Studies. April 2020.
[57] National Academies of Sciences, Engineering, and Medicine, Committee on Human-System Integration Research Topics for the 711th Human, Performance Wing of the Air Force Research Laboratory, Board on Human-Systems Integration, Division of Behavioral and Social Sciences and Education, and Board on Human-Systems Integration. *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, D.C.: National Academies Press. 2022.
[58] Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 'Deep Reinforcement Learning from Human Preferences'. *ArXiv:1706.03741 [Cs, Stat]*. 13 July 2017.
[59] Veeramachaneni, Kalyan, Ignacio Arnaldo, Alfredo Cuesta-Infante, Vamsi Korrapati, Costas Bassias, and Ke Li. 'AI2 : Training a Big Data Machine to Defend'. p.13. 2016.

Human-machine interfaces may either contribute to address the predictability problem – if interfaces allow for better understanding, overseeing, and control of the AI system – or exacerbate it, for example, if interfaces make an AI system less understandable or visible to the humans interacting with it. Interface limitations, like transparency requirements leading to information overload or not enabling memory of past interactions with users, can increase the cognitive overhead of the human operators or reduced situational awareness.[60]

4.2 Training

Training and experience-building programs can help operators to calibrate their expectations of an AI system, and form more accurate representations of the system's general behaviour to overcome interface issues. They can also lead to the improvement of interfaces altogether through iterative design based on end-user feedback and user stories. These training programmes require novel concepts, methods, and standards especially when considering HMT-AI.[61]

To a large extent, HMT-AI literature has built on structures developed for more traditional HMT with lower levels of automation. It has also built on an understanding of human teams in which successful coordination depends on agents' abilities to "share representations, to

[60] Paleja, Rohan, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 'The Utility of Explainable AI in Ad Hoc Human-Machine Teaming'. In *Advances in Neural Information Processing Systems*, 34. pp.610–623. Curran Associates, Inc. 2021.

[61] Laird, John, Charan Ranganath, and Samuel Gershman. 'Future Directions in Human Machine Teaming Workshop', 2019; Lavin, Alexander, Hector Zenil, Brooks Paige, David Krakauer, Justin Gottschlich, Tim Mattson, Anima Anandkumar, et al. 'Simulation Intelligence: Towards a New Generation of Scientific Methods'. *ArXiv:2112.03235 [Cs]*. 6 December 2021; National Academies of Sciences, Engineering, and Medicine, Committee on Human-System Integration Research Topics for the 711th Human, Performance Wing of the Air Force Research Laboratory, Board on Human-Systems Integration, Division of Behavioral and Social Sciences and Education, and Board on Human-Systems Integration. *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, D.C.: National Academies Press. 2022.

predict other agents' actions, and to integrate the effects of these action predictions".[62] However, aspects and dynamics of these teams do not map perfectly to the characteristics of HMT-AI. From assigning dynamic roles and objectives in new contexts of deployment, to developing "shared mental models" (or representations) and trust, to the assignment of responsibility, HMT-AI require different and new approaches, which are currently underexplored.[63,64] Consider for example the utility of explainable AI (xAI) techniques, like decision trees, that can provide humans with insight into the AI agent's behaviour policies and limitations to enhance the situational awareness of the human operator and improve the development of shared mental models. Recent research suggests that the advantage of using xAI techniques to improve "team fluency" in HMT-AI can vary greatly depending on the team composition and levels of domain expertise of human teammates, potentially leading to performance degradation for experts [77, p. 6].

HMT-AI should involve regular training exercises that introduce uncertainties and perturbations, to help both humans and artificial agents construct well-rounded representations of each other's decision-making criteria and capabilities, as well as team-specific conventions.[65] For example, studies on trust in emergency guide robots[66] have highlighted the potential benefit for humans in HMT-AI to experience wrong behaviour from

---

[62] Paleja, Rohan, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 'The Utility of Explainable AI in Ad Hoc Human-Machine Teaming'. In *Advances in Neural Information Processing Systems*, 34. pp.610–623. Curran Associates, Inc. 2021; Sebanz, N, H Bekkering, and G Knoblich. 'Joint Action: Bodies and Minds Moving Together'. *Trends in Cognitive Sciences* 10, no. 2. pp.70-76. February 2006.

[63] McNeese, Nathan J., Beau G. Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 'Who/What Is My Teammate? Team Composition Considerations in Human-AI Teaming'. *ArXiv:2105.11000 [Cs]*. 23 May 2021.'; O'Neill, Thomas, Nathan McNeese, Amy Barron, and Beau Schelble. 'Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature'. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. 22 October 2020.

[64] Consider for example decision tree explanations of AI agents' hierarchical policies for action or inferring teammates' actions Paleja et al., 'The Utility of Explainable AI in Ad Hoc Human-Machine Teaming'.

[65] Niu, Yaru, Rohan Paleja, and Matthew Gombolay. 'Multi-Agent Graph-Attention Communication and Teaming', 2021. 10; Shih, Andy, Arjun Sawhney, Jovana Kondic, Stefano Ermon, and Dorsa Sadigh. 'On the Critical Role of Conventions in Adaptive Human-AI Collaboration'. *ArXiv:2104.02871 [Cs]*. 6 April 2021.

[66] Robinette, Paul, Ayanna M. Howard, and Alan R. Wagner. 'Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations'. *IEEE Transactions on Human-Machine Systems* 47, no. 4. pp.425-436. August 2017.

the robot prior to any use in real situations, so that the human can gain some awareness of and adapt to the machine's imperfections. At the same time, improved mental models developed by human agents during training sessions can be used to improve the performance of collaborative artificial agents or the interface that facilitates communication between agents, creating a feedback loop.[67] This approach has been suggested for*, inter alia,* HMT-AI in real time strategy games,[68] military war-gaming,[69] autonomous flight teaming,[70] and cybersecurity.[71] It is equally relevant in the context of intelligence analysis.

Security agencies in both the UK and the US[72] have stressed that developing effective training programs tailored to HMT unpredictability will require: building new simulation environments; figuring out when to train for perturbation/adaption versus standardisation; better machine models of humans and alignment mechanisms; and the reduction of over-trust. Further

[67] Klamm, J., C. Dominguez, B. Yost, P. McDermott, and M. Lenox. 'Partnering with Technology: The Importance of Human Machine Teaming in Future MDC2 Systems'. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 11006. pp.259–66. SPIE. 2019.

[68] Anderson, Andrew, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 'Mental Models of Mere Mortals with Explanations of Reinforcement Learning'. *ACM Transactions on Interactive Intelligent Systems* 10, no. 2. pp.1-37. 30 June 2020.

[69] Schwartz, Peter J., Daniel V. O'Neill, Meghan E. Bentz, Adam Brown, Brian S. Doyle, Olivia C. Liepa, Robert Lawrence, and Richard D. Hull. 'AI-Enabled Wargaming in the Military Decision Making Process'. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, 11413. pp.118–134. SPIE. 2020.

[70] Tossell, Chad, Boyoung Kim, Bianca Donadio, Ewart de Visser, Ryan Holec, and Elizabeth Phillips. 'Appropriately Representing Military Tasks for Human-Machine Teaming Research'. pp.245–65. 2020.

[71] Buchanan, Ben, and Andrew Imbrie. *The New Fire: War, Peace, and Democracy in the Age of AI*. 2022; Ding, Wen, Sonwoo Kim, Daniel Xu, and Inki Kim. 'Can Intelligent Agent Improve Human-Machine Team Performance Under Cyberattacks?' In *IHSI*. 2019; Gomez, Steven R., Vincent Mancuso, and Diane Staheli. 'Considerations for Human-Machine Teaming in Cybersecurity'. In *Augmented Cognition*, edited by Dylan D. Schmorrow and Cali M. Fidopiastis, 11580. pp.153–68. Lecture Notes in Computer Science. Cham: Springer International Publishing. 2019.

[72] National Academies of Sciences, Engineering, and Medicine, Committee on Human-System Integration Research Topics for the 711th Human, Performance Wing of the Air Force Research Laboratory, Board on Human-Systems Integration, Division of Behavioral and Social Sciences and Education, and Board on Human-Systems Integration. *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, D.C.: National Academies Press. 2022; Laird, John, Charan Ranganath, and Samuel Gershman. 'Future Directions in Human Machine Teaming Workshop'. 2019.

research is needed to understand how to achieve these results in practice. This leads us to the following two recommendations.

> **Recommendation 1.** Government research funding should be allocated to develop public-private collaborations and longitudinal studies on HMT-AI. This research should focus on old and new models for decision-making in HMT-AI to assess the impact of team conventions building and training on performance and control measures. Focus should be drawn on defining new training protocols for HMT-AI specific dynamics, and on accelerating the development of risk management standards and HMT-AI performance assessments.

> **Recommendation 2.** A dedicated certification scheme for HMT-AI should be established, to promote industry consensus on the design requirements and evaluation of AI systems designed for HMT-AI. Generalising between tasks, effective communication, performance consistency, and adapting to new teammates should all be included within such a certification scheme. Building on under-developed ISO standards,[73] this certification scheme should also extend to the traceability of processes and decision accountability as well as auditing mechanisms to evaluate levels of trust in HMT-AI. This is necessary to disincentivise over-trust and complacent attitudes in HMT-AI that maintain or amplify the predictability problem.

Understanding the nature of trust and its implications for the predictability problem is the goal of the next section.

[73] ISO. 'ISO/IEC 38507:2022 Information Technology — Governance of IT — Governance Implications of the Use of Artificial Intelligence by Organizations'. ISO. 2022; ISO/IEC DIS 23894 Information Technology — Artificial Intelligence — Guidance on Risk Management'. ISO. 2022.

4.3 Trust and Trustworthiness

Trust is a facilitator of interactions among members of a system or team, whether human agents, artificial agents, or a combination of both, as in HMT-AI.[74] There are many definitions of trust. Often, these are domain-specific and focus on ethical, psychological, economic, societal, and security aspects of trust. However, when considering the nature of trust from a philosophical point of view, it becomes clear that occurrences of trust are related to, and affect, pre-existing relations. Examples include purchasing, negotiation, communication, and delegation.[75] In this sense, trust is not a relation itself but something that qualifies relations. It is a *property of relations*.

As a property of relations, trust changes the way relations occur by making it convenient for the agent who decides to trust (the trustor) to engage in the relation. It does so by minimising their efforts and commitments (e.g., time and resources) for the achievement of a given goal. The trustor saves efforts and commitments in two ways. First, they can delegate an action necessary to achieve their goal. Thus, they avoid performing the action themselves, because they can count on the trustee to do it. Second, the trustor can decide not to supervise (or to supervise less) the trustee's performance. Delegation without (or with limited) supervision is the mark of an ideal relation of trust.[76] It is because of this facilitating role that trust is crucial for any system to work. Without trust, delegation would be much more problematic as it would require a constantly high level of supervision. And this, in turn, would encroach upon the distribution of tasks necessary for most systems to function. Imagine not trusting the GP, the

[74] Primiero, Giuseppe, and Mariarosaria Taddeo. 'A Modal Type Theory for Formalizing Trusted Communications'. *Journal of Applied Logic* 10, no. 1. pp.92-114. March 2012.
[75] Taddeo, Mariarosaria. 'Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust'. *Minds and Machines* 20, no. 2. pp.243-257. 15 June 2010; Taddeo, Mariarosaria. 'An Information-based Solution for the Puzzle of Testimony and Trust'. *Social Epistemology* 24, no. 4. pp.295-299. October 2010.
[76] Taddeo, Mariarosaria. 'Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust'. *Minds and Machines* 20, no. 2. pp.243-257. 15 June 2010.

children's teacher, or the mechanic. This would imply spending a significant portion of time and resources monitoring the way they perform their tasks.

Trust comes in levels (Alice may trust Bob more than she trusts Charles, for example). Levels of trust can be assessed based on the level of supervision over the trustee performing a given task: the higher the level of supervision the lower the level of trust. The level of supervision depends (for rational, non-gullible, economic agents) on the level of trustworthiness of the trustee. This is assessed considering past performances of the trustee (their reputation) as well as the risks that the trustor will face if the trustee behaves differently from what is expected (risk/benefit analysis). When the probability that the trustee's expected behaviour will occur is either too low or not assessable, the risk is too high, and trust is unjustified. The predictability problem, along with the lack of transparency and vulnerability of AI systems,[77] makes it hard to evaluate whether the same system will continue to behave as expected in any given context. This impairs the assessment of its trustworthiness, both in terms of the reputation of the trustee and in terms of risks/benefit analysis.

Coming back to AI, HMT-AI, and the predictability problem, this is not to say that we should not trust the teams with high-impact decisions and tasks, especially when HMT-AI can perform them efficiently and efficaciously. On the contrary, delegation can and, in appropriate cases, *should* still occur. However, some forms of supervision are necessary to mitigate the risks linked to the predictability problem. The level of supervision should be proportionate to the level of trustworthiness of HMT-AI systems and the risks that unpredicted behaviour would entail (discussed further in sections 5.4 and 5.5). Policy strategies focused on eliciting users' trust without considering appropriate forms of supervision or monitoring in HMT-AI will fail to address this crucial issue, and exacerbate the risks linked to the predictability

[77] Taddeo, Mariarosaria, Tom McCutcheon, and Luciano Floridi. 'Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword'. *Nature Machine Intelligence* 1, no. 12. pp.557-560. December 2019.

problem. To avoid these risks, it is crucial to foster the right level of trust in AI systems for any given decision-making context.

4.4 Levels of Trust in HMT-AI

In an ideal scenario, the level of trust is proportionate to the trustworthiness of the trustee. Thus, over-trust occurs when the trustor has a high level of trust in a trustee whose trustworthiness is low. Similarly, under-trust happens when despite the high trustworthiness of the trustee, the trustor disproportionately oversees their behaviour. Human agents in HMT-AI are exposed to the risks of over-trusting the artificial agents with which they work (automation bias). For example, a 2016 study describes an experiment involving 42 volunteers in a simulated fire emergency with a robot guide tasked to lead them to safety.[78] Nearly 90% of the participants followed the robot blindly as it committed several fatal mistakes for which it never provided explanations nor warnings.

When occurring in high-stakes decision-making contexts, over-trust aggravates the risks of unpredictability and can lead to adverse outcomes. Over-trust generates *trust and forget* dynamics,[79] whereby the trustor has the highest level of trust in the trustee and does not supervise its performance, to the extent of overlooking its (potentially erroneous) actions, disregarding its capabilities and limits, and accepting uncritically its outcomes. Usually, these dynamics stop when something goes (badly) wrong and recalls our attention to the trustworthiness of the trustee.

[78] Robinette, Paul, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 'Overtrust of Robots in Emergency Evacuation Scenarios'. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. pp.101–108. Christchurch, New Zealand: IEEE. 2016.

[79] Taddeo, Mariarosaria. 'Trusting Digital Technologies Correctly'. *Minds and Machines* 27, no. 4. pp.565-568. 1 December 2017.

Over-trust can be a consequence of automation bias in AI – i.e. the tendency of human agents to over-rely on AI outcomes as these are perceived as more accurate or better than human-driven solutions.[80] As Struß argues, this bias and the risk of over-trust become more problematic as the complexity of an AI system increases (deep automation bias). This is because the artificial and the human agent in HMT-AI have different decision-making processes, and the human agents may be unable to scrutinise or understand how the artificial agent reaches its decisions, but still decide to rely on it due to the effects of the automation bias.[81]

When over-trust and *trust and forget* dynamics occur in HMT-AI involved in high-stakes decisions, they risk leading to negative consequences for the human-machine team, organisations or societies at large. In the security context, these dynamics coupled with the predictability problem could lead to severe risks for citizens' rights and real-world security risks. Avoiding these risks requires defining and developing ways to identify the right levels of trust within an HMT-AI and discourage over-trust. The training measures described in section 4.2 and recommendations 1 and 2 will help address this issue.

[80] Goddard, Kate, Abdul Roudsari, and Jeremy C Wyatt. 'Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators'. *Journal of the American Medical Informatics Association* 19, no. 1. pp.121-127. January 2012.

[81] It is worth noting that every application of AI will have different data models or ML approaches (from supervised learning to reinforcement learning), or sensors, and this will affect the level of automation embedded in the AI system. In turn, the level of automation in the system will affect the ability for human agent to scrutinise and understand its outputs.

5. Addressing the Predictability Problem with Good Governance

Taking stock of the analysis of the root causes of the predictability problem identified in the previous sections, we shall now analyse which of these causes are covered in relevant policy documents, and if so how and to what extent. In this section, we focus specifically on UK policies relevant to the national security community, but also consider EU and US policies in the same domain (Table 2).

Region	Type of document
UK policies	UK National AI Strategy [89]
	UK National Cyber Strategy [90]
	Policy report of the Government Communications Headquarters (GCHQ) [91]
	Policy report of the Royal United Services Institute (RUSI) [92]
	Policy report of the House of Lords Select Committee on AI (2019)
	Policy report of The Alan Turing Institute [94]
	Policy report of Chatham House [95]
	Policy report of the National Cyber Security Centre (NCSC) [96]
	Policy report of the Defence Science and Technology Laboratory (Dstl, 2020) [97]

EU policies	EU Ethics Guidelines on AI [98]
	The proposed EU AI Act [99]
	Policy report from the EU Joint Research Centre [100]
	Policy report from the EU Agency for the Operational Management of Large-Scale IT Systems [101]
	Policy report from the European Union Agency for Cybersecurity [102]
US policies	US Department of Defence Data Strategy [103]
	The final report from the US National Security Commission on AI (NSCAI, 2021)

Table 2. The list of policy documents analysed in this report and the related geographical origin.

When reviewing these policy documents, we identified four themes and a gap. These are: control, oversight, and value alignment; the resource boosting approach; the development of trustworthy AI; and the lack of focus on risk management measures to curtail the impact of the predictability problem. We analyse each of these in turn and identify points of improvement and recommendations, to inform future policy responses focused on addressing the predictability problem.

5.1 Control, Oversight and Value Alignment

Policy documents and proposals on balancing human and machine decisions often focus on control and oversight,[82] framed as a series of responses to the human inability to predict fully the behaviour of AI systems. For example, the Joint Research Centre (JRC) report identifies the complexity of AI models as a challenge to inspection and control by human operators.[83] The European Union Agency for the Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice (EU-LISA) proposes developing multiple tools that can be used for system failure detection and prediction, according to the availability and quality of data sets.[84] Other policies focus on defining control and oversight standards for testing and verification.[85] Some documents propose approaches such as securing "human-in-the-loop",[86] "human-on-the-loop"[87] and "human-in-command" deployment parameters.[88] Others call for the imposition of operational constraints on the system at the design phase. Often these documents mention human oversight[89] to refer to different modalities of human control at different stages of the AI lifecycle.

[82] Babuta, Alexander, Marion Oswald, and Ardi Janjeva. 'Artificial Intelligence and UK National Security: Policy Considerations'. Occasional Paper. London: Royal United Services Institute for Defence Studies. April 2020; GCHQ, 'GCHQ | Pioneering a New National Security: The Ethics of Artificial Intelligence'. 2021.

[83] JRC. *Robustness and Explainability of Artificial Intelligence: From Technical to Policy Solutions.* LU: Publications Office. 2020.

[84] EU LISA, and Aleksandrs Cepilovs. *Artificial Intelligence in the Operational Management of Large-Scale IT Systems: Research and Technology Monitoring Report : Perspectives for Eu LISA*. LU: Publications Office of the European Union. 2020.

[85] McAleese, Madison Jones. 'Will AI Prediction Technology Impact National Security?' Pacific Council on International Policy. 7 August 2018.

[86] GCHQ, 'GCHQ | Pioneering a New National Security: The Ethics of Artificial Intelligence'. 2021.

[87] Stumborg, Michael, and Becky Roh. 'Dimensions of Autonomous Decision-Making'. Mark. 2021.

[88] European Commission, and Directorate-General for Communications Networks, Content and Technology. *Ethics Guidelines for Trustworthy AI*. LU: Publications Office of the European Union. 2019.

[89] European Commission, and Directorate-General for Communications Networks, Content and Technology. *Ethics Guidelines for Trustworthy AI*. LU: Publications Office of the European Union. 2019; McKendrick, Kathleen. 'Artificial Intelligence Prediction and Counterterrorism'. Chatham House, The Royal Institute of International Affairs. 2019.

Value alignment – the need for AI systems to be coherent with values and responsibilities of humans – is another aspect covered in this context.[90] Proposals of standards and certification procedures to foster an ecosystem where there is alignment between AI goals and human values also abound.[91] This derives from concerns around unpredictability in the face of increasing levels of automation[92] or complexity of systems, like the black box problem.[93] More automation increases the stakes of unpredictability, demanding more stringent alignment mechanisms.

Whether they refer to control, human oversight or value alignment, the proposals outlined in these documents assume a clear separation between humans and the systems with which they work. However, the relationship between the two is arguably more organic and bi-directional than what most documents assume. As discussed previously, AI systems are increasingly embedded in HMT.[94] This reframes the scope of the problem, extending it to the realm of interactions between human agents and their environment. Building policies with HMT-AI at the centre, rather than at the periphery, would lead to more effective governance and design mechanisms.

By focusing on HMT-AI, national security policy documents would characterise AI systems as parts of a system broadly understood, one consisting of human as well as AI agents with respect to their relevant dimensions for the predictability problem. In contexts where human-

[90] JRC. *Robustness and Explainability of Artificial Intelligence: From Technical to Policy Solutions.* LU: Publications Office. 2020.

[91] JRC. *Robustness and Explainability of Artificial Intelligence: From Technical to Policy Solutions.* LU: Publications Office. 2020; McKendrick, Kathleen. 'Artificial Intelligence Prediction and Counterterrorism'. Chatham House, The Royal Institute of International Affairs. 2019; Schwartz, Reva, Apostol Vassilev, Kristen K. Greene, Lori Perine, Andrew Burt, and Patrick Hall. 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence'. 15 March 2022.

[92] Boardman, Michael, and Fiona Butcher. 'An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenges of Achieving It'. STO-MP-IST-178. 2019; Defence Science and Technology Laboratory. 'Building Blocks for AI and Autonomy: A Dstl Biscuit Book'. GOV.UK. 2020.

[93] Holland Michel, 'The Black Box, Unlocked | UNIDIR'. 2020.

[94] Stumborg, Michael, and Becky Roh. 'Dimensions of Autonomous Decision-Making'. Mark. 2021.

machine interfaces are used to tackle security threats, e.g. in cybersecurity, this would amount to framing human behaviour (e.g. human intuition in the detection of threats or their reactions to an AI system's inferences) as an integral aspect of the predictability problem, rather than a solution to an AI system's unpredictability. While prior elements of unpredictability were attributed mostly to machines, this framing would allow for the mapping and integration of human and environmental factors that can be sources of unpredictability in human-machine interactions. These factors include policy changes or cognitive aspects of decision-making within the set of risks that need to be monitored and managed.

This shift in approach would stand as a bridge between the maximal and the minimal definition of the predictability problem, as it allows us to frame problems that were likely to be attributed to the former (e.g. technical problems in AI design and development) and those likely to be attributed to the latter (e.g. human over-trust of an AI system after deployment) as transversal and interacting issues. At the same time, integrating concerns around control and oversight, such as autonomy or complexity of AI systems, as part of the wider "team" dynamics can help in the process of matching strengths and weaknesses of human and machine agents, which is a key aim of HMT-AI.[95]

Shifting the policy focus towards HMT-AI would lead to more holistic and realistic policy strategies to mitigate predictability-related risks of HMT-AI failures, risks to citizens' rights and security, as well as risks of reputational damage to institutions. This leads us to the following recommendation:

[95] Stumborg, Michael, and Becky Roh. 'Dimensions of Autonomous Decision-Making'. Mark, 2021.

Recommendation 3. Policy responses to the predictability problem in the national security domain should focus on governing HMT-AI teams, rather than AI systems alone.

5.2 The Resource Boosting Approach: The Risk of Overlooking Predictability Trade-offs

Several policy documents and proposals suggest responding to risks related to the predictability problem through a resource boosting approach: they embrace a logic of more data,[96] more coordination and collaborations,[97] skills upgrading,[98] and more funding.[99] If not coupled with measures considering context of deployment and societal impact, this approach is problematic and risks narrowing the policy focus on techno-centric and cumulative solutions, and overlooking ethical and social risks.

Better predictability of the outcomes of both AI systems and HMT-AI are a necessary requirement when considering whether to deploy AI for national security purposes. This is, however, only one of the criteria that should be factored into the risk/benefit analysis driving the decision to use, or not use, AI technology in this domain. For instance, in the national security context, some applications of AI have the potential to impact significantly on human rights, most notably Article 8 of the European Convention on Human Rights (ECHR), the right to respect for one's private and family life. As any activity that has the potential to impact on such rights must be assessed as both necessary and proportionate in the interests of national

[96] McKendrick, Kathleen. 'Artificial Intelligence Prediction and Counterterrorism'. Chatham House, The Royal Institute of International Affairs. 2019.
[97] Desouza, Gregory S. Dawson and Kevin C. 'How the U.S. Can Dominate in the Race to National AI Supremacy'. *Brookings* (blog). 3 February 2022; HM Government, 'National Cyber Strategy'. 2022. 35; Babuta, Alexander, Marion Oswald, and Ardi Janjeva. 'Artificial Intelligence and UK National Security: Policy Considerations'. Occasional Paper. London: Royal United Services Institute for Defence Studies. April 2020.
[98] GCHQ, 'GCHQ | Pioneering a New National Security: The Ethics of Artificial Intelligence'. 2021; Gorman, Christopher. 'Recent Developments in AI and National Security: What You Need to Know'. Lawfare. 3 March 2022; HM Government, 'National Cyber Strategy'. 2022.
[99] HM Government, 'National AI Strategy'. 2021; HM Government, 'National Cyber Strategy'. 2022.

security,[100] so decisions as to whether to use AI need to be made by weighing up the potential benefit offered, balanced against the potential risk to individual rights.

Existing frameworks regarding the respect of fundamental rights as well as risk assessment in national security are already in place, and The European Convention on Human Rights, the United Nations International Covenant on Civil and Political Rights, and the United Nations Siracusa Principles offer guidance when specifying necessity, proportionality and scientific validity as the criteria to restrict any measure that might impinge on human rights.[101] With regards to risk assessment, cost-benefit analysis (CBA) and necessity and proportionality considerations are already in place.[102] However, further guidance is required regarding adapting these frameworks to the AI predictability problem, focussing on which aspects are more (or less) relevant and how these calculations should be made in practice for HMT-AI.

It is also problematic that, where guidance is offered for this decision, the interpretation and application of this guidance is, at times, expected to be self-administered. As it is likely that oversight and scrutiny in this area may not be public, it is even more important that the CBA is conducted by independent bodies, which should be enabled and supported to develop an objective, in-depth assessment and should be accountable to the public for their assessment.

> **Recommendation 4.** CBA of HMT-AI in the national security domain should include an assessment of the predictability of AI systems and of the related ethical risks along the technical and operational dimensions. To facilitate coherent assessment across security agencies, a standard scale to assess predictability of AI systems

[100] European Parliament, Council of the European Union. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), 201 OJ L §. 2002; Gov UK, Investigatory Powers Act 2016.

[101] Morley, Jessica, Josh Cowls, Mariarosaria Taddeo, and Luciano Floridi. 'Ethical Guidelines for SARS-CoV-2 Digital Tracking and Tracing Systems'. SSRN Scholarly Paper. Rochester, NY. 22 April 2020.

[102] Gov UK, Investigatory Powers Act 2016.

should be defined, where the choice of using (or not) AI should be justified on this scale with respect to a contextual CBA as well as the consideration of public attitudes towards the risks and the benefits involved. The definition of this scale should be within the remit of an independent third-party actor, i.e., a different public office than the one deploying the HMT-AI. Ideally, the CBA should also be run by an independent body, or at the very least it should be independently scrutinised.

The resource boosting approach often leads to calls for more data and more complex AI models to mitigate the predictability problem, under the assumption that these would lead to greater accuracy. Several documents consider data as a strategic asset and they exhort to make data secure, trustworthy, and interoperable.[103] The European Union Agency for Cybersecurity (ENISA) talks about data augmentation techniques when training datasets are too small.[104] In terms of more complex AI systems, the UK National Cyber Strategy proposes to "scale up and develop law enforcement technical capabilities",[105] which can be deployed against threats. In the same vein, while urging caution, the UK Royal United Services Institute considers the use of augmented intelligence techniques,[106] such as cognitive automation of human sensory processes with NLP and audio-visual analysis and behavioural analytics.[107] Others urge the use of more complex models such as neural networks, combined with symbolic reasoning – e.g. neuro-symbolic AI[108] – to increase the accuracy of results without excessively sacrificing explainability.

[103] Norquist, David L. 'DOD Data Strategy'. 2020. 16.
[104] ENISA, 'Artificial Intelligence Cybersecurity Challenges'. Report/Study. 2020.
[105] HM Government, 'National Cyber Strategy'. 2021.106.
[106] Babuta, Alexander, Marion Oswald, and Ardi Janjeva. 'Artificial Intelligence and UK National Security: Policy Considerations'. Occasional Paper. London: Royal United Services Institute for Defence Studies. April 2020'; GCHQ, 'GCHQ | Pioneering a New National Security: The Ethics of Artificial Intelligence'. 2021.
[107] Babuta, Alexander, Marion Oswald, and Ardi Janjeva. 'Artificial Intelligence and UK National Security: Policy Considerations'. Occasional Paper. London: Royal United Services Institute for Defence Studies. April 2020.
[108] European Commission, AI Act Proposal. 2021.

However, as discussed previously, the sharing and constant accumulation of large datasets to improve technical predictability is also a contributor to increased unpredictability, as data sources and labelling become harder to validate and large datasets become attractive targets for malicious actors.[109] Additionally, if more data is added incrementally, the AI system will require updating and re-tuning. As the UK National Cybersecurity Centre states, 'updates to the system will change the performance of the tool, which may cause it to become unpredictable' (NCSC, 2019, p.6). More data and more complex models might decrease concerns around predictability at a superficial level, while affecting other system properties or mechanisms that may compromise predictability more indirectly. As stressed at the beginning of this report, there is a minimal and a maximal understanding of predictability. More data and complexity can work in favour of one and yet against the other. For example, while choosing a more complex AI model might extend its capacity to respond to a set of different scenarios and decrease unpredictability in terms of its ability to respond reliably to external factors (as described in the maximal interpretation of the problem), it might also increase unpredictability in terms of our capacity to understand a system's internal behaviour after deployment and thus predict it (as described in the minimal interpretation of the problem).

> **Recommendation 5.** Rather than "more" or "less" predictability, policy proposals should focus on predictability trade-offs, making clear which aspect of the predictability problem specific proposals aim to tackle and in which way, as well as which aspects they risk exacerbating, and which mitigating measures will be put in place. Policies should recognise that predictability is a multi-dimensional concept,

[109] Ananny, Mike, and Kate Crawford. 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability'. *New Media & Society* 20, no. 3. pp.973-989. March 2018; ENISA, 'Artificial Intelligence Cybersecurity Challenges'. Report/Study. 2020.

where gains in predictability on one level can come at the expense of losses on another.

We do not mean to suggest that predictability is a relative concept, but rather a multi-dimensional one, and policy proposals for more data and complexity should be assessed against their differential impact on these multiple dimensions.

5.3 Trustworthiness: Unjustified Trust in the Face of the Predictability Problem

Trustworthiness of HMT-AI is a property that needs to be assessed over time, across multiple metrics and agents. It is problematic when trustworthiness is presented as tantamount to predictability, misleading readers to think that 'if it's predictable then it's trustworthy', or when predictability is described as a criterion for the use of AI alongside trustworthiness, as if predictability and trustworthiness were independent. This confuses any attempt to formalize the relation between trustworthiness and its assessment criteria, and risks misleading the assessment of trustworthiness of AI.

To this end, it does not help that trustworthiness is paired to a host of elements in policy documents that can affect predictability. Most documents and policy papers discuss robustness,[110] reliability,[111] safety,[112] and security[113] as elements of trustworthiness but do not consider their impact on the predictability of AI systems. These elements are related to predictability to the extent that they may lead to unpredictable outcomes. For example,

[110] ENISA, 'Artificial Intelligence Cybersecurity Challenges'. Report/Study. 2020; European Commission, and Directorate-General for Communications Networks, Content and Technology. *Ethics Guidelines for Trustworthy AI*. LU: Publications Office of the European Union. 2019; JRC. *Robustness and Explainability of Artificial Intelligence: From Technical to Policy Solutions.* LU: Publications Office. 2020.

[111] Holland Michel, 'The Black Box, Unlocked | UNIDIR'. 2020; Taddeo, Mariarosaria, Tom McCutcheon, and Luciano Floridi. 'Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword'. *Nature Machine Intelligence* 1, no. 12. pp.557-560. December 2019.

[112] ENISA, 'Artificial Intelligence Cybersecurity Challenges'. Report/Study. 2020.

[113] ENISA, 'Artificial Intelligence Cybersecurity Challenges'. Report/Study. 2020.

problems of security are not the same as problems of predictability. However, a system whose security is weak or compromised is more exposed to attacks which, in turn, make the system more likely to show unpredictable behaviours.

In the same vein, the EU Ethics Guidelines on AI present Trustworthy AI as a product of lawful, ethical and robust properties.[114] Other policy documents list trustworthy AI as a distinct property alongside other elements such as reliability and robustness. For example, in relation to cybersecurity of AI, ENISA states that it is crucial that a system is "trustworthy, reliable and robust" (ENISA, 2020a, p.30). JRC states that the European Commission has committed itself to a "trustworthy and secure use of AI*"* (JRC, 2020, p.1).

This approach inflates the concept of trustworthy AI and may lead to over-trusting a system once it has been labelled "trustworthy". Consider, for example, a possible use of behavioural analytics for counterterrorism; over-trusting a system could entail using unreliable inferences about alleged threats, possibly leading to a disproportionate escalation and undue breaches of individual rights.

Policy documents have attempted to operationalize the values and elements cited above. In this respect, some documents go as far as to provide check-lists,[115] self-assessment tools[116] or guidance for the use of AI tools.[117] While these are important, the lists often consist of series of questions and lack the specification of priority or hierarchy among elements and the definition of processes and assessment criteria.

[114] European Commission, and Directorate-General for Communications Networks, Content and Technology. *Ethics Guidelines for Trustworthy AI*. LU: Publications Office of the European Union. 2019.
[115] European Commission, and Directorate-General for Communications Networks, Content and Technology. *Ethics Guidelines for Trustworthy AI*. LU: Publications Office of the European Union. 2019.
[116] UKSA, 'Ethics Self-Assessment Tool'. UK Statistics Authority, 2019.
[117] NCSC, 'Intelligent Security Tools'. 2019.

If left unaddressed, this approach might turn "trustworthiness" into a blue-washing label[118] rather than a solid basis for deciding whether to rely on an AI system. For example, a company might declare a certain product trustworthy after a partial fulfilment of the checklist, after self-administered guidance or by fiat. For trustworthiness to become a reliable criterion driving the decision to use an AI system, the gap from "what" to "how"[119] needs to be filled through conceptually sound, operationally feasible, and accountable solutions.

> **Recommendation 6.** Policies on the problem of AI predictability in national security should address the link between trustworthiness and unpredictability, both at a formal and operational level. For example, AI systems should be given an amendable predictability score, which should be included in the assessment of the trustworthiness of the system. The trustworthiness of an AI system should include the CBA to assess the risks that unwanted behaviour may pose in different contexts of deployment.

5.4 A Notable Absence: Risk Thresholds for Unpredictable AI and the Predictability of Risks

Often, policy documents treat unpredictability as an element that can increase risks. However, they do not categorise these risks. To this end, we argue that it is important to recognize that some unpredictable scenarios or behaviours are riskier than others. At the same time, some risks are more predictable than others, leading to a 'meta-level' of overall risk. Let us focus first on the risks related to the predictability problem: the risks of unpredictability.

---

[118] Floridi, Luciano. 'Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical'. *Philosophy & Technology* 32, no. 2. pp.185-193. 1 June 2019.

[119] Babuta, Alexander, Marion Oswald, and Ardi Janjeva. 'Artificial Intelligence and UK National Security: Policy Considerations'. Occasional Paper. London: Royal United Services Institute for Defence Studies. April 2020; Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices'. *Science and Engineering Ethics* 26, no. 4. pp.2141-2168. 1 August 2020.

Once a certain unpredicted scenario occurs, the consequences can be highly problematic, and when national security is involved, potentially devastating. Consider, for example, the "automatic mode" (management by exception) of the US Patriot Missile system which generated false targets and engaged in friendly fire as a result.[120] As mentioned previously, relevant policy papers discuss catastrophic consequences of AI in terms of fatal errors and failures.[121] At the same time, it is possible that the realization of an unpredicted scenario can be harmless, or even lead to new positive outcomes and discoveries. This may depend, for example, on whether the sector itself is high-risk (e.g. national security, defence, healthcare, transport) and whether the intended use involves high-risk decisions or actions (e.g. injury, death, restriction of liberty, significant material/immaterial damage).[122]

Policy proposals that focus on risk related to AI-based solutions or HMT-AI[123] do not provide criteria to define risk-thresholds for unpredictable AI. That is, identifying a level of unpredictability that makes the deployment of a given solution too dangerous. Criteria for this assessment should encompass technical as well as ethical, legal, and social considerations around what counts as "risky" or "more/less risky" in the face of unpredictability.

Focusing now on the meta-level of risk (plainly, on how risk itself might suffer from a predictability problem), policy papers rarely elaborate on how some risks are more predictable than others, with some being not predictable at all.[124] Admittedly, identifying all possible risks in a context and considering their likelihood is not logically impossible, but it is

[120] Sheridan, Thomas B. 'Human Supervisory Control Of Automation'. In *Handbook Of Human Factors And Ergonomics*, edited by Gavriel Salvendy and Waldemar Karwowski, 1st ed. pp.736–60. Wiley. 2021.
[121] Hawley, Dr John K. 'Automation and the Patriot Air and Missile Defense System'. 2017. 18
[122] European Commission, AI Act Proposal. 2021.
[123] ENISA, 'Artificial Intelligence Cybersecurity Challenges'. Report/Study. 2020; European Commission, AI Act Proposal. 2021; HM Government, 'National Cyber Strategy'. 2022.
[124] Cameron, David, Great Britain, and Cabinet Office. *A Strong Britain in an Age of Uncertainty: The National Security Strategy*. London: TSO. 2010; Stumborg, Michael, and Becky Roh. 'Dimensions of Autonomous Decision-Making'. Mark. 2021; HM Government, 'National Cyber Strategy'. 2022.

unfeasible. The number of variables and their possible interactions is exorbitantly large, and this makes any assessment intractable.

When considering unpredictability of AI systems or HMT-AI, what is relevant is the type of risks (more or less predictable), rather than the sole number of risks involved. A distinction can be made among risks in terms of:

- known knowns;
- known unknowns; and
- unknown unknowns.

Assuming that we can predict what we know, the above list could be translated into a categorization of risks on a scale from more to less predictable. Examples of the first would be reasonably expectable failures, such as the decay of a certain system after a long time and the need for maintenance efforts. Examples of the second would be threats, such as cyberattacks, whose possibility we are aware of but whose occurrence is difficult to predict. An example of unknown unknowns are so-called black-swan events:[125] rare and unpredictable outlier events which we can make sense of only retrospectively. A classic example is the crash of the US housing market during the 2008 financial crisis.

When applied to the predictability problem, this categorization can help to inform policy responses by strategizing risk-mitigating approaches, prioritising risks based on their predictability or their impact. For example, it could make more sense to build robustness first against risks we are aware of but cannot predict, and then focus efforts on addressing risks that we are aware of and can predict. In the domain of national security, it is reasonable to build first resilience and robustness towards potentially catastrophic AI system failures that

[125] Taleb, Nassim Nicholas. *The Black Swan: The Impact of the Highly Improbable*. Random House Publishing Group. 2007.

cannot be predicted, and then meet the maintenance requirements of systems to retain stability and prevent failure.

> **Recommendation 7.** Risk thresholds should be established for unpredictable AI which map the severity of risks around unpredictable behaviour to their own level of predictability (e.g., division into known knowns, known unknowns, etc.). These thresholds will in turn inform the development of risk management processes, allowing risks to be prioritised based on their predictability and their impact.

The analysis proposed in this section shows that a coherent risk assessment framework is needed to identify and mitigate risks associated with the predictability problem. The framework should offer criteria to assess risks across different domains. International standards – like ISO/IEC SC42 and the draft AI Risk Management Framework by the US National Institute of Standards and Technology[126] – are being developed to fill that gap. In the UK, the ALARP (as low as reasonably practicable) engineering principle may offer a foundation for developing an AI-specific risk assessment framework. We turn to this principle and its relevance for the predictability problem in the next section.

5.5 An ALARP-based Framework to Assess the Risk of Unpredictable AI

The ALARP principle is similar to the "as low as reasonably achievable" principle in the nuclear industry[127] and has been widely applied in safety-critical industries,[128] its practical

[126] NIST, 'AI Risk Management Framework: Initial Draft'. 2022.
[127] Baybutt, Paul. 'The ALARP Principle in Process Safety'. *Process Safety Progress* 33, no. 1. pp.36-40. March 2014.
[128] Gai, Wen-mei, Yan Du, and Yun-feng Deng. 'Evacuation Risk Assessment of Regional Evacuation for Major Accidents and Its Application in Emergency Planning: A Case Study'. *Safety Science* 106. pp.203-218. July 2018;

implementation discussed in research and policy[129] and "adopted within engineering good practice as a proportionate approach to safety risk management".[130] It has also been used by the UK Health and Safety Executive (HSE).

The ALARP principle operates in line with the precautionary principle, and relies on duty-holders' adoption of good industry practices for controlling risks related to hazardous situations.[131] It sets goals for duty-holders rather than imposing prescriptive regimes, and requires professional judgment rather than formal calculations. It has four prescriptions designed to help practitioners identify the externalities of multidimensional problems:[132]

1. Identify hazards on their proposed facility;
2. Analyse the hazard, and determine the risk arising from that hazard (as specified in the previous recommendations);
3. Take actions to reduce the risk from the hazard, by either designing it out or providing appropriate controls or protection;
4. Demonstrate to the regulators in each case that the residual risks to workers and public from the facility are tolerable and ALARP.[133]

Langdalen, Henrik, Eirik Bjorheim Abrahamsen, and Jon Tømmerås Selvik. 'On the Importance of Systems Thinking When Using the ALARP Principle for Risk Management'. *Reliability Engineering & System Safety* 204. December 2020. 107222; Morley, Bob. 'Best Practicable Means (BPM) and as Low as Reasonably Practicable (ALARP) in Action at Sellafield'. *Journal of Radiological Protection* 24, no. 1. pp.29-40. 1 March 2004.

[129] Ale, B.J.M., D.N.D. Hartford, and D. Slater. 'ALARP and CBA All in the Same Game'. *Safety Science* 76. pp,90-100. July 2015; Guikema, S.D., and T. Aven. 'Is ALARP Applicable to the Management of Terrorist Risks?' *Reliability Engineering & System Safety* 95, no. 8. pp.823-827. August 2010; Jones-Lee, M., and T. Aven. 'ALARP—What Does It Really Mean?' *Reliability Engineering & System Safety* 96, no. 8. pp.877-882. August 2011.

[130] Department of Transport, 'Principles and Guidelines for the Spaceflight Regulator in Assessing ALARP and Acceptable Risk', 2021.

[131] HSE, *Reducing Risks, Protecting People*. Sudbury: HSE Books. 2001.

[132] Hurst, John, Jenna McIntyre, Yoshikazu Tamauchi, Hiroshi Kinuhata, and Takashi Kodama. 'A Summary of the 'ALARP' Principle and Associated Thinking'. *Journal of Nuclear Science and Technology* 56, no. 2. pp.241-253. 1 February 2019.

[133] Hurst, John, Jenna McIntyre, Yoshikazu Tamauchi, Hiroshi Kinuhata, and Takashi Kodama. 'A Summary of the 'ALARP' Principle and Associated Thinking'. *Journal of Nuclear Science and Technology* 56, no. 2. p.244. 1 February 2019.

Many distinctions made in the ALARP principle overlap with the engineering and operational aspects of AI systems, specifically with respect to its approach to different types of uncertainties (e.g., knowledge uncertainty, modelling uncertainty and limited predictability). For this reason, it is relevant with respect to AI systems to identify, assess, and accept/refuse related risks in an appropriate way. This leads us to the following recommendation:

> **Recommendation 8.** An ALARP-based framework should be developed to assess the risks of unpredictable AI and HMT-AI, and establish the maximum acceptable degree of unpredictability for any given context. This framework should include:
>
> - A quantitative assessment of the level of predictability of a given AI system and HMT-AI;
> - An assessment of the traceability of the design, development, and/or procurement steps leading to deployment of the AI system;
> - An assessment of the conditions of deployment, e.g., HMT-AI, level of training of operators (or HMT-AI members), level of transparency of the interface, level of human control over the AI system;
> - A cost-benefit analysis of the potential risks and intended benefits of deploying the system (as per Recommendation 4);
> - An analysis of hypothetical scenarios to consider how exposure to risk or the effectiveness of mitigating measures may vary with context of deployment;
> - Protocols for human overriding of the system and redress mechanisms.

Four advantages follow from the ALARP-based approach. First, its widespread adoption in UK industry makes it a good candidate to facilitate skill transfers across industries and collaboration with fields that are already familiar with the principle. Second, the predictability problem can be mapped against the tolerability of risk (TOR) framework to ensure more

appropriate predictability thresholds depending on contexts of deployment.[134] Third, the ALARP principle could help policymakers and practitioners to determine which AI systems should not be used in certain contexts due to the identification of unacceptable risks. Even where the use of AI systems would not be completely ruled out, such risks may suggest that only much more predictable AI systems should be used. For example, an organisation may opt to deploy deterministic systems in high-risk contexts to increase the predictability of AI-driven actions. Fourth, the principle fosters an economic deployment of AI systems. It emphasises the concept of gross disproportion, that is an unreasonable imbalance between the severity of the risks and the costs of mitigating them. For example, both the severity of risks and the cost of taking measures against them will on average be lower for rules-based AI systems than deep neural networks, or for offline models rather than online ones. Yet, in the recently proposed "Algorithmic transparency data standard", the UK's Central Digital and Data Office[135] urges government departments to complete an evaluation for every algorithmic tool – not distinguishing between levels of complexity and predictability. The cost of this measure can be considered grossly disproportionate for smaller teams using numerous common algorithmic tools with rules-based behaviour.

It is important to stress that while the proposed ALARP-based framework would offer useful guidance, alone it would not be sufficient to identify and mitigate the risks posed by the predictability problem and AI-based HMT for national security purposes. To this end, it is crucial that all eight recommendations provided in this report are implemented, more research is developed, and a stronger focus on the ethical, legal, and social implications of using HMT-AI in this domain is fostered as part of wider institutional approaches to AI development and use.

[134] HSE, *Reducing Risks, Protecting People*. Sudbury: HSE Books. 2001.

[135] Central Digital and Data Office and Office for Artificial Intelligence, 'Assessing If Artificial Intelligence Is the Right Solution'. GOV.UK. 2019.

6. Conclusion

By focusing on the predictability problem, in both its maximal and minimal interpretation, this report has provided a new level of analysis to consider risks related to learning capabilities, lack of transparency, low robustness, complexity, and threats to AI systems. We believe that this enables a better approach to dealing with these risks, because it identifies a common thread – the predictability problem – among them.

The findings of this report should be considered preliminary. Further work needs to focus on a better grasp of the root causes of the predictability problem, and identifying design criteria for HMT-AI that would enable human agents to leverage the potential of AI agents, while fostering human autonomy. In the same vein, future work should focus on what governance approaches should be implemented to incentivise mechanisms that limit the negative impact of the predictability problem, including assessing the predictability of AI systems, and ensure accountability and redress mechanisms for error or unintended outcomes. This is a particularly pressing need when considering the use of AI systems in the national security domain while respecting democratic values and fundamental rights. As the UK government's special report on HMT stresses, it is not an overestimation to state that unpredicted behaviour could have catastrophic consequences,[136] not least in terms of public trust in, and reputation of, the agencies deploying them.

Deploying AI systems without a proper foundation of design and policy strategies around what makes a system unpredictable in the first place risks automating unpredictability. This would make the use of AI systems in high-stake contexts ethically and operationally problematic. Further interdisciplinary research on the predictability problem and its ethical,

[136] Hawley, 'Automation and the Patriot Air and Missile Defense System'. 2017. 18; MoD, 'Human–Machine Teaming. Joint Concept Note 1/18.' 2018.

legal, and social implications is essential to ensure that future national security uses of AI are effective, ethically sound, and proportionate.

Appendix – Glossary

Accuracy: The degree to which the Machine Learning (ML) model is correctly capturing a relationship that exists within training data.[137]

Accountability: Determinations of accountability in the AI context are related to expectations for the responsible party in the event that a risky outcome is realised. Individual human operators and their organisations should be answerable and held accountable for the outcomes of AI systems.[138]

Algorithm: A set of step-by-step instructions. Computer algorithms vary in complexity and produce different approaches to learning. The following algorithms can be applied to almost any data problem:

- Linear Regression, Logistic Regression, Decision Trees, k-nearest neighbours (Supervised learning);
- K-means, neural networks, a priori algorithms (Unsupervised learning);
- Markov Decision process (Reinforcement learning).

Artificial Intelligence (AI): "The use of digital technology to create systems capable of performing tasks commonly thought to require intelligence. AI generally:

- involves machines using statistics to find patterns in large amounts of data
- is the ability to perform repetitive tasks with data without the need for constant human guidance."[139]

[137] NIST, 'AI Risk Management Framework: Initial Draft'. 2022.
[138] NIST, 'AI Risk Management Framework: Initial Draft'. 2022.
[139] Central Digital and Data Office and Office for Artificial Intelligence, 'Assessing If Artificial Intelligence Is the Right Solution'. GOV.UK. 2019.

Black box: "A description of some deep learning systems. They take an input and provide an output, but the calculations that occur in between are not easy for humans to interpret".[140]

Data Quality: The extent to which data is fit for purpose. Data quality measurements entail various dimensions based on the context of use. Some commonly used dimensions are completeness, accuracy, uniqueness, timeliness, consistency and validity.[141]

Deep learning: "How a neural network with multiple layers becomes sensitive to progressively more abstract patterns. In parsing a photo, layers might respond first to edges, then paws, then dogs".[142]

Explainability: The ability for a user to understand how the model works.[143]

Hazard: The potential for harm arising from an intrinsic property or disposition of something to cause detriment.[144]

Human Machine Teams (HMT): "Mutually beneficial coordination of humans and machine intelligence systems (which include AI and ML algorithms and models, various data and information flows, hardware including compute architecture and sensor arrays, etc.)".[145]

Machine learning: A system that was designed to learn patterns and associations in input data. A machine learning algorithm is designed to specify the creation and modification of a mathematical model of the relationship between input and output data. Once the model is

---

[140] Hutson, Matthew. 'AI Glossary: Artificial Intelligence, in so Many Words'. *Science* 357, no. 6346. 19-19. 7 July 2017.
[141] Loshin, David. 'Data Quality and MDM'. In *Master Data Management*. pp.87–103. Elsevier. 2009.
[142] Hutson, Matthew. 'AI Glossary: Artificial Intelligence, in so Many Words'. *Science* 357, no. 6346. 19-19. 7 July 2017.
[143] NIST, 'AI Risk Management Framework: Initial Draft'. 2022.
[144] HSE, *Reducing Risks, Protecting People*. Sudbury: HSE Books, 2001.
[145] Lavin, Alexander, Hector Zenil, Brooks Paige, David Krakauer, Justin Gottschlich, Tim Mattson, Anima Anandkumar, et al. 'Simulation Intelligence: Towards a New Generation of Scientific Methods'. *ArXiv:2112.03235 [Cs]*. 6 December 2021.

trained it is able to do many tasks including classification, prediction and recommendations.[146]

Neural Network: "A highly abstracted and simplified model of the human brain used in machine learning. A set of units receives pieces of an input (pixels in a photo, say), performs simple computations on them, and passes them on to the next layer of units. The final layer represents the answer".[147]

Predictability (technical): Assessed as the degree of consistency between a system's past and current behaviours and its future ones.[148]

Reinforcement learning: "A type of machine learning in which the algorithm learns by acting toward an abstract goal, such as "earn a high video game score" or "manage a factory efficiently." During training, each effort is evaluated based on its contribution toward the goal".[149]

Reliability: Indicates whether a model consistently generates the same results, within the bounds of acceptable statistical error.[150]

Risk: The chance that someone or something that is valued will be affected adversely in a stipulated way by the hazard.[151]

Robustness: "A measure of model sensitivity, indicating whether the model has minimum sensitivity to variations in uncontrollable factors. Measures of robustness might range from

[146] Hutson, Matthew. 'AI Glossary: Artificial Intelligence, in so Many Words'. *Science* 357, no. 6346. 19-19. 7 July 2017.
[147] Hutson, Matthew. 'AI Glossary: Artificial Intelligence, in so Many Words'. *Science* 357, no. 6346. 19-19. 7 July 2017.
[148] Holland Michel, 'The Black Box, Unlocked | UNIDIR'. 2020.
[149] Hutson, Matthew. 'AI Glossary: Artificial Intelligence, in so Many Words'. *Science* 357, no. 6346. 19-19. 7 July 2017.
[150] NIST, 'AI Risk Management Framework: Initial Draft'. 2022.
[151] HSE, *Reducing Risks, Protecting People*. Sudbury: HSE Books, 2001.

sensitivity of a model's outputs to small changes in its inputs, but might also include error measurements on novel datasets."[152]

Structured Data: Highly organised, standardised, categorised and easily understood data, often represented in terms of rows and columns (tabular data).

Supervised Learning: "A type of machine learning in which the algorithm compares its outputs with the correct outputs during training. In unsupervised learning, the algorithm merely looks for patterns in a set of data."[153]

Trust: A form of delegation of a task with no (or low level of) supervision of how the delegated task is performed. Its occurrence makes it convenient for the agent who decides to trust (the trustor) to engage in the relation, as the trustor saves time and resources to achieve a given goal by delegating related tasks. The lack, or low level of, supervision implies some risks for the trustor, should the trustee behave differently than what expected.[154]

Technical Debt: Long-term software issues stemming from forgoing best practices at development stage in favour of easier and quicker solutions.

Uncertainty: "A state of knowledge in which, although the factors influencing the issue are identified, the likelihood of any adverse effects or the effects themselves cannot be precisely described.

Unstructured data: Data that does not have predefined standards for processing and management and requires specialised tooling, infrastructure and expertise compared to

[152] NIST, 'AI Risk Management Framework: Initial Draft'. 10. 2022.
[153] Hutson, Matthew. 'AI Glossary: Artificial Intelligence, in so Many Words'. *Science* 357, no. 6346. 19-19. 7 July 2017..
[154] Taddeo, Mariarosaria. 'Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust'. *Minds and Machines* 20, no. 2. pp.243-257. 15 June 2010; Taddeo, Mariarosaria, Tom McCutcheon, and Luciano Floridi. 'Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword'. *Nature Machine Intelligence* 1, no. 12. pp.557-560. December 2019.

structured data. Examples of unstructured data include images, text, IoT sensor data, audio, and video.

Versioning: The practice of keep track of all the versions of a software or AI model, their performances, and hyperparameters.

About the Authors

Mariarosaria Taddeo,[1, 2] Marta Ziosi, Andreas Tsamados,[1] Luca Gilli,[3] Shalini Kurapati[3]

[1] Oxford Internet Institute, University of Oxford, UK

[2] Alan Turing Institute, London, UK

[3] Clearbox AI, Turin, Italy

Mariarosaria Taddeo is Associate Professor and Senior Research Fellow at the Oxford Internet Institute, University of Oxford, and is Defence Science and Technology Laboratory (Dstl) Ethics Fellow at the Alan Turing Institute. She also serves on the AI Ethics Advisory Committee of the UK Ministry of Defence.

Marta Ziosi is a DPhil at the Oxford Internet Institute and a member of the Digital Governance Group at the University of Oxford. Marta is also the co-founder and the chair of "AI for People", a non-profit research association that promotes the use of AI starting from human and societal needs.

Andreas Tsamados is a DPhil student in Information, Communication and the Social Sciences at the Oxford Internet Institute, University of Oxford. His research focuses on the ethical, environmental and social implications of digital technologies, with a strong focus on AI.

Luca Gilli is the co-founder and CTO of Clearbox AI. He holds a PhD in Applied Physics (Uncertainty Quantification) from TU Delft. He is also a member of the United Nations/ITU working group on Data and AI Solution Assessment Methods.

Shalini Kurapati is the co-founder and CEO of Clearbox AI. She holds a PhD in Technology, Policy and Management from TU Delft. She specialises in data management, data privacy and ethics. She is a Certified Informational Privacy Professional/Europe (CIPP/E).

CETaS

Centre for Emerging Technology and Security

RESEARCH REPORT