AI for Security Topic Guide Issue 1.0.0

Matilda Rhode | Airbus

EDITOR
Awais Rashid | University of Bristol

REVIEWERS
Elisa Bertino | Purdue University
Daisuke Mashima | Advanced Digital Sciences Center
Guillermo Suarez-Tangil | IMDEA Networks Institute

COPYRIGHT

The CyBOK project would like to understand how the CyBOK is being used and its uptake. The project would like organisations using, or intending to use, CyBOK for the purposes of education, training, course development, professional development etc. to contact it at **contact@cybok.org** to let the project know how they are using CyBOK.

1 INTRODUCTION

Cyber security, like other industries, has seen an explosion in the use of artificial intelligence (AI) and machine learning (ML) technologies in recent years to help automate tasks. Data-driven approaches in general can draw patterns from vast volumes of data far quicker than humans can. This short introduction summarises the state of AI for security at the time of writing and highlights some of the considerations to guide whether it is an appropriate approach for a given problem, common pitfalls to avoid, and human-AI ecosystems.

AI is challenged by several open research areas including lack of transparency, robustness to concept drift, and the security of AI systems themselves. The first two issues are addressed in this topic guide and the third in Security and Privacy for AI Knowledge Guide [25] .

This topic guide is aimed at those looking to build and/or procure AI solutions in relation to cyber security applications. Terms in bold are defined in the Glossary (B).

The next section introduces AI and ML (2) before asking 'Why use AI for security?' (3), with a brief overview of the potential benefits (3.1) and challenges (3.2). Applications are addressed in Section 4 organised under the NIST Cyber security Framework pillars [39].

Both builders and procurers may be interested in the sections on common pitfalls (5), evaluating AI (6), and lifecycle management (7). For data scientists and model builders, Section 8 discusses various incoming certifications (8.3) and common governance themes of privacy (8.4), robustness and concept drift (8.5), bias and explainability (8.6), feature engineering (8.1) and algorithm selection (8.2). Procurers may be more interested in Section 9, which recommends evaluation methodologies and questions to ask of opaque solutions.

2 AI AND MACHINE LEARNING

Cyber security products increasingly feature the terms 'AI', 'Machine Learning' (ML) and 'Deep Learning'. The terms 'Statistics' and 'Data Science' appear less frequently but may strongly underpin ML technologies. Machine Learning can be described as a sub-field of both AI and of data science.

The terminology of these overlapping scientific fields is sometimes loosely applied; in particular 'AI', a broader field, is frequently used to describe technologies using ML and statistics. This section will briefly introduce AI and ML with reference to their use for cyber security applications.

2.1 Artificial Intelligence

Artificial intelligence (AI) is a wide field of interdisciplinary research defined by its goals rather than methods. It concerns *intelligent agents* where intelligence may be the emulation of *human intelligence* or *rational intelligence* [92]. Cyber security applications are usually best-served by rational agents rather than mimicking human behaviour e.g. for detecting and mitigating attacks. However, human intelligence is also of interest, e.g., for automated adversary modelling [67] or to simulate cyber professionals and user behaviour [70] for sandboxing, training and risk modelling.

AI addresses a number of problems: capturing and storing information (perception, knowledge representation), processing information (planning, learning, decision making), communication, and action. Agents do not need to address all of these to be considered part of AI.

AI may be further described as "narrow" or "general". **Narrow AI** addresses a limited and specific problem (e.g., playing arcade games, driving a car, defending a computer network) whereas **Artificial general intelligence (AGI)** aims to tackle any problem it is presented with, in the way a human could. AGI does not currently exist. At the time of writing there has been discussion as to whether significant advances towards more general intelligence have been made in natural language processing (NLP) since several organisations publicly released conversational AI large language model (LLM) interfaces [73, 81] capable of answering questions on a wide range of topics, summarising text, and writing code [107]. For cyber security, preliminary research has suggested the use of these technologies for finding vulnerabilities, generating secure technologies such as hardware [68] and even to write malware [17]. LLMs rely heavily on a subfield of AI: machine learning.

2.2 Machine Learning

Machine learning infers rules to map data to objectives according to an algorithm. The algorithm may try to solve a number of different problem types such as classification, clustering, anomaly detection or regression. ML typically relies on large volumes of data in order to recognise patterns, e.g., between malicious and benign network traffic or between individuals' biometric data for authentication. However, some ML research focusses on low-data learning, e.g., zero-shot learning problems whereby a model attempts to address a problem for which it has not been trained.

Machine learning algorithms 'learn' by iteratively adjusting a set of parameters during a training phase. Upon satisfying some criteria, the model parameters are frozen and the model is tested. The model may remain in this frozen state forever. Alternatively, it may be updated through periodic training or continuous learning whilst deployed. **Neural networks** are just one category of ML algorithm, loosely inspired by the human brain, comprising networks of nodes connected by weighted pathways. **Deep learning** is a subfield of neural networks using multiple sequentially stacked layers of nodes. These algorithms have grown famous for their use in breakthroughs in computer vision, NLP and other tasks [14].

Model training is an optimisation problem, seeking to maximise positive feedback such as the error rate between the ground truth and the model's predictions. Three (and a half) broad categories of machine learning models can be described by the feedback given to the model during training:

[1] Supervised learning uses ground-truth labelled datasets. The cost of labelling data is often high because it requires expert evaluation, but model accuracy is also high [52] thus reducing the human effort required to interpret how useful the outputs are.

Security application examples: distinguishing benign and malicious emails, software, network traffic, classifying malware into families (see CyBoK Malware & Attack Technologies Knowledge Area (KA) Section 4.2.2 for ML-based analytics[59]); predicting the financial cost of a given attack (see Risk Management & Governance Knowledge AreaSection 3 for predicting cost [21]).

[2] Unsupervised learning does not use labelled data for training, though some algorithms do require metadata such as the percentage of anomalous samples in the training data. Labelled data may only be needed for model evaluation, if at all. However, the outputs of unsupervised models often require human interpretation; for example, anomaly detection for malicious user behaviour uses unsupervised learning but human analysts must verify whether an anomaly is simply unusual or constitutes a threat to security.

Security application examples: anomalous network traffic, user or process activity detection (see Security Operations and Information Management Knowledge AreaSections 3.2, 3.3 and 3.4 for more on anomaly detection including the use of machine learning and section 7 on human misuse detection); clustering attacks for attribution. (see Malware & Attack Technologies Knowledge AreaSection 4.2.2 for more on attribution [59]).

Semi-supervised learning should reap some of the benefits of supervised learning without the high cost of generating a huge number of labels by hand. This may either by using a small corpus of labelled examples together with a larger group of unlabelled samples. The labelling strategy for the small corpus can take a number of forms such as making use of existing labelled data or by dynamically querying a human expert as in **active learning** [108]. Alternatively, a two-step process with no human labels can be employed using techniques such as **self-learning or self-training** [9, 108].

Security application examples: attack detection (see [59] section 4.2.2 as above and [116]), network monitoring (see [89] section 4.2).

[3] Reinforcement learning (RL) takes feedback from its environment, the feedback is often framed as rewards. RL generates training data through exploratory actions, observations and feedback. There is no labelling cost for RL since the feedback (labels) are directly gathered from the environment through this exploration. Two significant challenges in RL are how to specify (i) the environment (often a simulation) and (ii) the rewards [92] so that the model learns useful strategies rather than trivial shortcuts.

Security application examples: automated red-teaming, automated cyber defence (blue-teaming) (see [32] Section 6 for defensive incident response and intelligence).

3 WHY AI?

A wide range of cyber security products now advertise the inclusion of AI technologies with the vast majority of these products being data-driven and reliant on machine learning and/or statistics. The remainder of this guide will use the term 'AI' to encompass both the machine learning as well as the smaller set of non-ML cyber security applications.

It's worth examining why AI has become so popular in cyber security products. The simple answer is that it has become popular for all kinds of products. There are two key drivers of this phenomenon: (i) hardware breakthroughs enabling fast-training of deep learning models, with cloud infrastructures making the hardware increasingly accessible and (ii) very large data reserves (big data) that enable sharing such data, also benefitting from cloud-networking for storage and data sharing. Deep learning has accelerated technological leaps in self-driving cars, virtual assistants and game-playing for humans. These headline-grabbing breakthroughs drive the use AI in other fields. In some cases, researchers transform cyber security problems into, e.g., computer vision problems [61] in order to take advantage of research from another domain.

What are the benefits of using this technology for cyber security? Is it useful or simply fashionable? We might start by asking 'What is the difference between machine learning and other software?' Non-ML software uses hand-coded rules to transform inputs. Machine learning uses inductive reasoning to develop a model that will transform inputs to a given set of outputs without explicit programming. Therefore, for typical ML programs no human *needs* to know the input-output-mappings that make up the software (model), this can be both an advantage and a disadvantage.

3.1 Potential benefits

(Partial) automation of tasks ML technologies can automate tasks fully or partially so that security professionals can concentrate on other priorities.

Automatic retraining Technology and the threat landscape both evolve quickly, meaning that hand-written rules require frequent updating. ML can update these rules automatically and instantly by learning from new data.

Big data analysis Computer networks, users, and attackers all generate data; too much for a (team of) humans to sift through. ML can draw out relevant patterns efficiently from terabytes of data. See [32] Section 2 for an overview of data sources used in security monitoring.

Harnesses (latent) information Additional data that has not been of interest to human analysts may contain valuable insights that are worth uncovering.

3.2 Challenges

Reliance on large datasets Large volumes of data may not be readily available, e.g. low-volume high-impact attacks crafted for a specific target.

Simulated data may also be created using emulated or **digital twin** environments [33]. Further research looks at using **zero-shot learning** [111] which uses learning from other domains and transfers it to address a new problem, often reliant on **transfer learning** [117].

Cost of labelling data For training supervised models and validating most models, it is likely that a human will have to provide ground truth labels, which negates some of the benefits of automation.

Mitigations: The cost of labelling may be mitigated using techniques such as **semi-supervised learning** or **active learning** [27, 96, 118]. Crowdsourcing labels using online platforms may be possible for non-expert tasks [112] and/or trusted label sharing may be conducted between experts.

Lack of benchmark datasets Due to the rapid evolution, sensitive nature, and commercial incentives surrounding security data, there is a lack of data-sharing in the cyber security community. This makes it difficult to benchmark experimental research and products without conducting some data collection and testing oneself.

Mitigations: For research benchmarking, more public datasets are needed, but there is little motivation for companies or researchers to do so, with even testbed-generated data sometimes kept private due to concerns that the activity traces might leak sensitive information. Somewhat perversely, collecting private data may be beneficial for product benchmarking since publicly available data may already have been identified as malicious using other slow and/or non-automated techniques, thus artificially inflating the apparent performance of the AI/ML solution. Public or private datasets collected over time will help with benchmarking and robustness testing (see Section 8.5).

Data privacy ML security products may rely on (attack) data from their customers to keep models relevant, but this is not always acceptable if customers prefer to keep data private for security, business or employee and customer privacy, for example.

Mitigations: Privacy-preserving technologies such as **federated learning**, **differential privacy** and **homomorphic encryption** may improve privacy; see Section 8.4 for more on model and data privacy and the CyBoK Privacy & Online Rights Knowledge Area Section 1.1 for a deeper discussion of data privacy and relevant technologies [103].

Infrastructure cost Large datasets have large storage requirements and many ML models use power-hungry hardware, particularly during model training.

Mitigations: Preferring low-data and low-resource models could reduce this cost. By buying an external ML product, costs may be mitigated through economies of scale. Pre-trained models or generic cloud-based ML models accessed through an API may be more economical, as long as other security requirements are also satisfied.

Opacity, robustness, and security of ML Many data-driven AI and ML technologies are considered opaque since the input-output-mappings are not explicitly programmed and sometimes highly complex. This can make it difficult to diagnose *why* a model gave a particular output, whether the model has learned spurious correlations, whether it will fail to classify new data well, or stand up to adversarial attacks.

Mitigations: These problems may be jointly addressed by hardening models and reducing opacity through explainable AI (XAI) practices. At the time of writing these are still maturing research fields with demonstrably successful approaches still to be tested in the field. See sections 8.5 (robustness) and 8.6 (bias and XAI) and for adversarial attacks: Security and Privacy for AI Knowledge Guide [25] .

Incoming regulation New governance on the horizon points to many of the opacity, robustness and security problems above, but given the gap between problem and reliable solutions deployed technologies, may require retrofitting to meet a variety of new legislation. See Law & Regulation CyBoK Knowledge Area section 7.5.3 Affirmative defences including footnotes for compliance challenges [24].

Mitigations: Some new legislation is forecasted by primers giving clues to the likely requirements and is usually aligned with existing software-security principles. See Section 8.3.

4 CYBER SECURITY APPLICATIONS

AI and ML technologies have been applied to a wide range of cyber security problems, with a heavy focus on detection and response technologies. This section discusses applications of ML and AI using the NIST Cyber Security Framework [39] pillars for ease of reference to particular cyber security activities. Established and emergent applications are then cross-referenced with CyBoK Knowledge Areas in Section 4.6. In order to keep this topic guide to reasonable length, some of the less mature applications are brief, readers should refer to the citations for more examples and context.

4.1 Identify

Vulnerability discovery, especially for software vulnerabilities (rather than network, human or business operation vulnerabilities) has been conducted with ML. For source code, recent work has looked at using large-language models to assist the detection of bugs and vulnerabilities in code prior to compilation [5]. For complied code, use of ML includes prediction of vulnerabilities based on summative code metrics, anomaly detection in code patterns, and vulnerable code-matching [40] and the volume of expected vulnerabilities [60, 57]. Some research [29] indicates that ML models resulted in reduced false positive rates by comparison with rule-based approaches; but ML models are limited in their ability to recommend the type of vulnerability or a method of remediation [40]. AI and ML may be used for vulnerability discovery in hardware by replicating behaviour of devices within a model [109]. Research also points to the identification of *human* vulnerabilities that may increase chances of cyber risky behaviour [16].

Vulnerability assessment, such as vulnerability criticality may also be predicted using ML models using various data including social media activity [101] as well as metadata from vulnerability databases [58]. Open source data-driven projects such as the Exploit Prediction Scoring System which aims to improve on CVSS using quantitative methods may be of interest [50].

Automated red-teaming (with digital twins) has been tested for military contexts for some decades [3]. These exercises are highly dependent on the environment in which they are trained. Often, the environment is simulated, including **digital twins**, because either the real environment cannot be taken offline for any period, is too expensive to clone, or provides data too slowly (especially for machine learning). Various AI techniques such as planning [67] and more recently ML, especially RL have been used for this, with a number of public challenges currently active [1, 102]. These approaches are limited by the environment in which they learn. For instance, malicious agents have a limited set of actions [35]; a real attacker may have more actions at their disposal (e.g., social engineering) and/or the context may be more complex (e.g., multiple attacks at once, skeleton defence team during holiday season). Additionally, the merits and demerits of using simulated environments are independent of AI and ML; but for ML in particular, lots of data is needed hence a simulated environment is usually essential for the learning phase therefore it is worth considering the fidelity and usefulness of any simulation environment. The advantage of AI/ML approaches is that they may discover new successful attack pathways from the huge set of possible combinations. These pathways are typically limited to known vulnerabilities and attack techniques since the model is usually specified with a finite **action space**.

Governance and compliance in cyber security may use NLP techniques in order to distill relevant data [99] or to check that written policies comply with the law [6]. But the use of these

NLP tools is limited in a similar way to other automated legal applications (e.g., resolution of parking tickets [98]) in that the models often need human oversight and verification. The 'Identify' and 'Protect' functions could be linked in future using automated controls driven by legal documents. However, this requires not only further development of NLP technologies but a sufficient **action space** for an automated agent to implement required changes.

4.2 Protect

Access control list insights [64] and auto-generation of suggested controls from user stories [46] as well as other approaches have been explored. Almost all research to date acknowledges limitations in the ability of ML to create perfect lists, therefore human oversight is required. ML may speed up the process but does not yet constitute a replacement. Some research [75] has proposed the use of reinforcement learning to update access control policies between IoT devices but it relies on negative feedback in order to reduce trust/confidence in an object therefore allowing some unwanted interactions to take place first. See [4] for an overview of the use of AI/ML in identity and access management.

Authentication using biometric indicators is reliant on ML technology. However, ML is also capable of generative tasks and underpins **deepfake** threats by which faces, voices and other data are synthesised and may represent a threat to some biometric authentication technologies. Behavioural biometrics [51] allow for continuous authentication rather than one-shot, and may be conducted using tremor detection, gait analysis or touchstrokes [97]. Behavioural biometrics may be more difficult to fake as long as this data is not typically available for scraping on the Internet; unlike images and video footage via (social) media.

4.3 Detect

There is a wealth of data running through each device and network being used at any given time. As such, attack detection based on continuous monitoring is a mature market and field of research. Anomaly detection, classification (malicious vs. benign or specific types of attack or activity), clustering similar activity and attack attribution can all be (partially) tackled using ML techniques.

Intrusion detection systems (IDS) analyse network traffic in order to detect malicious activity. ML has used all kinds of network data including individual packets, network flows and machine reputation data in order to detect attacks [20, 94] and these are built into so-called 'next-generation firewalls'. For research, due to the aforementioned challenges of collecting large labelled datasets, IDS datasets are often generated by researchers by carrying out a set number of known attacks. Models trained on these datasets should be deployed with care to ensure that other types of attack can also be identified, e.g., by retraining leaving one attack methodology out and then testing for detection [55].

Malware detection using ML is often argued to avoid the weaknesses of static, signature-based detection, has achieved high detection accuracy in many papers and is widely used in industry. Malware detection models are particularly susceptible to concept drift as malware authors respond to both automatic detection tools and new opportunities all the time, with thousands of new samples detected every day [106]. If using ML for malware detection, it is important to regularly assess the performance of the system to ensure that it is still achieving high detection accuracy (see section 6).

Phishing detection often uses a combination of NLP techniques and analysis on any attachments, embedded scripts, URLs, and so forth [30]. Similar to malware detection, the need for automated filtering has existed for sometime, therefore adversaries are already habituated to changing tactics in order to evade detection, making concept drift a key concern for the usefulness of the solution.

User entity behaviour analytics (UEBA) represents a fast-growing market for monitoring the wide range of data generated by users in order to mitigate incidental security violations as well as insider threats [115]. Many UEBA solutions use anomaly detection to profile 'normal' user behaviour and identify deviations from normality. The use of this technology can be sensitive as personal data is often used for model training and analysis; accordingly, these solutions may be subject to more governance than other tools.

Data aggregation and System Information Event Management Systems (SIEMs) allow for combining the outputs of all the models described above. This can help to build a clearer image more quickly of the nature and extent of a given attack or threat [62].

Attribution and attacker profiling can help organisations to understand the motivations of their adversaries and use this intelligence to mitigate future attacks. Data can include real-world information such as exploited vulnerabilities, software (malware) patterns as well as data generated by honeypots [53]. Some challenges for this application are (i) attackers may try and mask their identity (ii) techniques used by groups may change (concept drift), and (iii) there may not be enough data to train a sufficiently accurate ML model. Recent work explores using RL [48] and generative AI chatbots [65] to generate realistic data in a honeypot environment to draw out malicious behaviour for analysis.

4.4 Respond

Automated response products, at the time of writing, generally uses ML-driven detection to trigger rules-based responses since inexplicitly programmed responses may cause unforeseen business interruption. The likely first AI/ML use cases will seek to maintain business operation in the face of an attack, e.g., using software-defined networking [104] or adaptive control in a manufacturing environment [83]. ML/AI-driven automated blue teaming is an active research area though remains commercially unavailable as yet due to the aforementioned business risk and the same challenges that automated red teaming faces (see section 4.1).

Forensic investigation can be helped by ML to explore large volumes of data. As well as examining network, machine and user artefacts which could be clustered against existing attack data, a forensic investigator can use NLP technologies in order to partially automate the search for relevant textual information [10, 84]. ML can also be used to *detect* deepfakes, post-processing [74] or file tampering [37].

4.5 Recover

The recovery function is included here for completeness. Recovery includes restoration of assets damaged in a cyber incident [39] and as such, AI might be used to help determine asset criticality for backing up or implement automated recovery, in which case this becomes muddled with the response function. AI could be used to predict the last 'clean' state of a recovery image, and may already be being used to generate public statements following breaches that minimise reputational damage.

4.6 CyBoK Knowledge Area Cross-Reference

CyBoK KA	Existing applications	Emergent applications
Risk Management & Governance [21]	impact prediction	vulnerability discovery, vulnerability criticality assessment, automated compliance checking, public response generation, automated pentesting
Law & Regulation [24]	IP theft detection, fraud detection	automated cyber law infringement detection, forensic evidence validity probability
Human Factors [95]	UEBA, phishing detection	human risk prediction, cyber professional and user simulations
Privacy & Online Rights	UEBA, **deepfake** detection	
Malware & Attack Technologies [59]	malware detection, malware analysis	attribution
Adversarial behaviours [100]	attack graph analysis	attribution
Security Operations & Incident Management [32]	anomaly detection, data aggregation and analysis, alert correlation, false-positive reduction, intrusion detection, UEBA	attribution, automated red-teaming, automated penetration testing, automated response
Forensics [90]	analytics of computer artefacts, NLP analysis of text files, **deepfake** detection	forensic evidence validity probability
Authentication, Authorisation & Accountability [41]	biometric and behavioural authentication, audit log assessment	access control list generation
Software Security [82]	vulnerability discovery	automated penetration testing, automated red-teaming
Web & Mobile Security [36]	phishing detection	
Network Security [89]	intrusion detection	software defined networking, dynamic access control
Hardware Security [109]	vulnerability discovery	attack detection
Cyber Physical Security [22]	anomaly detection, intrusion detection, digital twins for simulated detection and response	automated response

Table 1: Cross reference of existing and emergent uses of AI and relevant sections in other CyBoK Knowledge Areas

5 COMMON PITFALLS

For building AI and ML models, there are a number of common pitfalls to avoid. In general, the key risk with AI models is that metrics are used in order to conveniently summarise performance in many different contexts, a single number is often (wrongly) used to compare models and masks their strengths, weaknesses and expected behaviour. It is a little like assessing people's mathematical abilities by using exam results; '95%' does not tell us whether someone is strong in algebra or geometry nor how difficult the exam was; furthermore we would never compare *Person A*'s results from *Exam1* with *Person B*'s results on *Exam 2*; though this is frequently done with AI and ML models [34].

Arp et al. [11] provide a useful and detailed overview of common pitfalls for machine learning models in security together with the frequency of occurrence in academic research and possible security implications. They highlight that the three most common mistakes (which are not acknowledged) are: (i) **sampling bias** which may lead the model to learn spurious correlations, (ii) **data snooping** [80] in which the model gains clairvoyant insights into future attack data through mismanagement of training and testing sets, and (iii) **lab-only evaluation** [86] whereby the model is only tested on data from one source, collected at one point in time; using the same source and time period from which the training data came.

Adding to this comprehensive paper we may consider the issue of **baselining** anomaly detection models. It may be difficult to obtain a baseline of 'normal' that does not contain any abnormal instances, especially if collected in a live environment.

Most models will only work reliably within some bounds. When these bounds are exceeded and the model enters a **failure mode**, it may be necessary to invoke a backup system, whether human or automated. Most products do this but few research papers mention a safety net of any kind.

6 EVALUATION

As highlighted above, bad ML models can seem good when poor evaluation is used. A strong evaluation methodology can be used to check if a model is ready to deploy, support lifecycle maintenance, and check for robustness. The evaluation methodology can be developed before the model is built; doing so can protect against confirmation bias, by which the researcher builds a model that reinforces their pre-existing views, e.g., that packet flow length is the best way to detect malicious network traffic.

6.1 Performance metrics

Each metric has its strengths and weaknesses. It is tempting to use a single metric such as accuracy, area under the curve (AUC) or F-Score, but each of these may hide weaknesses in the model [119]. Looking at a range of metrics including false positive and false negative rates will help to uncover weaknesses, especially if broken down into subgroups of samples (e.g., malware families, network attack types).

For models which require human interpretation, e.g., anomaly detection, (precision at) top-k and recall should be analysed. Precision at top-k gives the proportion of relevant instances ranked in the first K instances. Recall gives the proportion of relevant instances that generated alerts. The latter may be an intractible task if the data is unlabelled [12] but generating true anomalies and checking if they are captured, as well as comparing against other anomaly detection models or products can provide some additional evaluation. See Chapter 2 of [119] for clear explanations of key machine learning performance metrics.

For reinforcement learning it is useful to analyse the pathways taken by the model to check for spurious shortcuts [31].

6.2 Analysis of datasets

Analysis of datasets can also help to mitigate against bias and promote robustness. Many machine learning algorithms are optimised based on percentage of correct predictions. If the data used to measure this features a large number of one type of, e.g., attack, the model may peak at performance that identifies only this type of attack, ignoring those less-represented cases in the dataset. The same is true for evaluation, so it can be beneficial to breakdown performance by data type.

6.3 Testing with multiple datasets

Testing with multiple datasets with different underlying distributions can also highlight weaknesses [86]. K-fold evaluation should be avoided since it may lead to data snooping [11]. The datasets selected can be spread over time to test for temporal concept drift robustness, from different sources to test for contextual robustness and/or omit certain types of sample to test for the ability to generalise e.g. to unseen attacks. A robust evaluation methodology can be used once the model is deployed to check performance and that the model does not need retraining.

6.4 Lifecycle maintenance

Lifecycle maintenance differs for built and procured systems but the evaluation of the systems can be the same. In order to check that performance is not dropping rapidly, a plan can be implemented for regular evaluation. This will require some human labelling effort to validate, but with planning this could take advantage of existing efforts such as in-depth human investigations which already provide ground-truth for network traffic, malware etc. This can be conducted periodically, following near-misses or successful attacks or ideally both. The mitigation of poor performance may be to retrain the model or contact the product provider.

7 ECOSYSTEM FOR AI TECHNOLOGY

The ecosystem of a model includes the human and technological resources required to build, maintain, and use the model. This section focusses on the human costs around implementing AI solutions, which are sometimes advertised to automate and reduce human resource costs.

Consider the cost and security of any cloud infrastructures used to train or store models and data, even if only during the research phase since these tools often persist into deployment. For more on securing AI models, refer to Security and Privacy for AI Knowledge Guide [25] .

7.1 Humans in the Loop

Human-in-the-loop systems require human validation of AI/ML model outputs. Other systems may not require human oversight for each output but even if the desired outcome of an AI solution is to reduce the human resources required for a function, there may be new activities required to support the use of the tool, even if they are outsourced.

Table 2 outlines common roles that humans might play in interacting with AI/ML systems, these should be factored into the building or procurement of an AI/ML system.

Human role	**Typical occupation**	**Frequency**	**Typical technology or scenario**
Model-building	AI specialist / data scientist	Once	All and any
Model-updating	AI specialist / data scientist	Over lifetime	May be required to improve accuracy,
Labelling	domain expert	Over lifetime	Training and retraining supervised learning models; to evaluate all models over lifetime
Evaluation	Data scientist, pentester	Over lifetime	To validate model robustness and identify vulnerabilities prior to deploying; to demonstrate compliance; to evaluate model performance over lifetime
Using model output as one factor in decision making	Security professional, business analyst	As required	AI outputs may simply be one tool used by security professionals to inform decisions, e.g., risk management strategies
Real-time interpretation/validation of model output	Security professional	Each prediction of malicious activity	Anomaly detection requires investigation (typical SOC analyst activity), validate critical model decisions
Take-over in case of failure	Model user	Training required over lifetime	In cases where models fail or operate outside of reliable context it may be necessary for a human to take over in realtime; human-computer-interaction design critical to ensure smooth transition
Audit	External auditor	Once or more	With GDPR regulations requiring models and inputs to be saved over time and incoming AI certification legislation currently being proposed, it may be necessary for humans to investigate logs of AI systems, their historic inputs and outputs as well in the context of relevant events, therefore the cost of building XAI and/or transparency logs should be considered

Table 2: Common human interaction with AI models, likely frequency of interaction and typical context in which interaction is necessary

7.2 Labelling and label dynamics

The cost of labelling is highlighted in Sections 3.2 and 7 . Security experts' time is precious and the rate at which labelling can be performed may be quite limited, though even limited expert labelling has been shown to significantly improve model performance [66].

Some research has bootstrapped the labelling of data by using open-source pre-labelled datasets or free tools such as VirusTotal [49] to label malware samples. Bootstrap techniques may be subject to quality and dynamics issues. In the specific case of VirusTotal for malware labelling, Zhu et al. [121] point out that, over time, different antivirus engines change the assigned label (malicious or benign) for the same sample and that stability can be achieved by setting a minimum threshold of engines that classify a sample as malicious.

When labelling a dataset using bootstrapping techniques, consider whether the assigned labels match the goal of the AI being designed. In the case of malware, many programs are potentially-unwanted-programs (PUPs) and not strictly malicious. Depending on the use-case it may be beneficial to omit these samples or assign a third label. For forecasting the cost of different attack scenarios, average costs may not be relevant to your organisation (different jurisdiction and defensive measures may impact expected costs). This cost may change over time, therefore even if the data that the model is classifying is not changing, the distribution of the labels may evolve over time, representing another form of **concept drift** [105].

8 CONSIDERATIONS FOR IMPLEMENTING AI

This section discusses some of the key challenges that may face those seeking to build and/or deploy an AI solution for a cyber security use case, including high level model design choices, compliance, privacy, robustness, bias mitigation, and explaining the outputs of AI models.

8.1 Model Design - Feature Engineering

ML models use data to learn, but these models cannot ingest raw information from the world, it must be transformed into a format that the model can read. As well as making it machine-readable, irrelevant information can be omitted, data may be summarised, made noisy, or enriched. The process of transforming raw data into model inputs (features) is known as **feature engineering** [120]. The choice of data is critical to building a performant model both in terms of accuracy but also robustness [86], privacy, and even explainability.

Zheng and Casari [120] provide a strong introduction to feature engineering techniques. This section lists some possible pitfalls in the feature engineering process. It is tempting to collect as many features as possible and either let a feature selection algorithm automatically extract the best features and/or input them all to the machine learning model for it to determine which features to use, but this may lead to undesirable model attributes.

Ideally, in feature selection we keep relevant features and omit irrelevant ones, but often we do not know which ones are relevant and which are not. Including too many, or the wrong features could lead to a high accuracy during validation but may mask an overfitted model that will not generalise well to new data; see [45] for a comprehensive overview on overfitting due to feature selection. Therefore, pruning features to maintain only relevant ones (using expert domain knowledge) can be critical to avoiding spurious correlations. Unfortunately this can be non-trivial and statistical methods can give insights for features which may or may not be relevant. Transforming feature extraction and reduction methods should be used with caution (e.g., using principal component analysis [79] or autoencoders [47]) if explainability is required.

Feature engineering may also restructure data for use by a particular algorithm. This process can both remove relevant structural information, e.g., NLP analysis of phishing email text may only look at short sequences of words (n-grams). Conversely, spurrious structural data could be introduced by, e.g., visualising malware as an image, thus creating non-existent relationships between data based purely on the selected image width [15]. Sometimes the former is necessary to meet computational resource constraints. Thorough testing of different feature sets with a variety of test sources may uncover hidden weaknesses before they are deployed.

8.2 Model Design - Algorithm selection

Algorithm selection similarly should be conscious of the context and structure of data. Like feature engineering, algorithm choice is too large a subject for this topic guide. Typically the structure of data, computational resource, latency and explainability requirements help to narrow the choices [119].

When selecting an algorithm, understanding its weaknesses may expedite debugging and robustness testing. For example, convolutional neural networks, whilst very widely used with high accuracy to evaluate image data, are well-known to focus on subsections of images rather than the over-arching structure [93]. This could be a significant weakness in some contexts.

8.3 Certification and Compliance

At the time of writing, legislation for AI systems is being widely-proposed by national and international regulatory entities. These vary from sector to sector and primarily focus on a subset of 'critical applications' including those which handle personal data and those interacting with physical systems. Some of the key legal frameworks include the General Data Protection Regulation (EU) [28], the Data Protection Act 2018 (UK) [43], the Equality Act 2010 [44] with guidelines published elsewhere such as: NIST AI Risk Management Framework (US) [69], Ethical Norms for New Generation Artificial Intelligence (China) [72] [1]. Legislation and standards are expected before 2025 from the EU (AI Act), US (NIST standards for trustworthy AI), and Canada (Artificial Intelligence & Data Act), for which preliminary documents or directives [78, 69, 71] have already been published.

Recently, proposed AI regulations are being discussed with respect to generative AI and whether the incoming legislation will be sufficient for any additional risks posed by this technology. This indicates the challenge and importance of developing future-looking certification. These discussions may catalyse changes in the form of regulation and/or changes to the regulatory timeline.

Common themes of these legislative proposals are **privacy**; **robustness**, especially relating to reliability and safety; and **bias**, with a strong focus on explaining models and their outputs.

[1] English translation: https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/

8.4 Privacy

Privacy-preserving machine learning is the easiest to demonstrably satisfy of these three criteria; using mathematical expressions to describe specific criteria. Privacy may be required to protect training data, data used during inference or even the the model itself. This may be to protect individuals' personal data, companies' security data or intellectual property; though legislation is mostly concerned with the first of these. In some cases it is possible to reverse-engineer the training data of a model either to expose details of training data or verify membership of a certain sample in the training data [23]. See CyBoK Privacy and Online Rights Knowledge Area for more on data (Section 1.1) and metadata (Section 1.2) confidentiality during processing.

As well as implementing sufficient access control measures, privacy-preserving learning methods have been developed to mitigate against these attacks. Attacks on the training data could be mitigated by privacy-preserving learning methods including injecting noise during training as in differentially private stochastic gradient descent [2] or by using voting ensembles of models [77]. In order to keep inputs private but benefit from learning across a number of private entities, e.g., malware samples discovered by different customers of an antivirus or detect-and-respond product, techniques such as **federated learning** [26], which combines data from locally-trained models or ML with **homomorphic encryption** [8] may be used. For secure-sharing of outputs, **secure multi-party computation** can be used to share only data that meets certain conditions. Protecting the IP of a model means defending against model inversion attacks. Homomorphic encryption again may be used here, refer to Security and Privacy for AI Knowledge Guide [25] for more information on preventing model inversion and other attacks on ML models. Multiple privacy-preserving methods may be required to meet various criteria.

8.5 Robustness and Concept Drift

The robustness of ML models is often discussed but definitions of robustness are not widely agreed from a quantitative perspective [18]. Colloquially, we might say that a model is not robust because it does not maintain performance under new unseen data, but research typically refers to specific kinds of robustness, e.g., robustness against temporal concept drift or against adversarial samples specifically crafted to fool the model.

Lack of robustness in security applications could lead to poor performance detecting new threats or model weaknesses being actively exploited by attackers. This topic guide focusses on robustness to new threats, often called concept drift, see Security and Privacy for AI Knowledge Guide [25] for exploits crafted specifically for AI. Robustness against new threats is difficult to ensure because we cannot predict all future inputs (e.g., attacks) to the model and AI/ML is typically used on problems for which the input space is so large that we cannot rely on traditional software verification methods.

Improving robustness against concept drift requires a means to detect concept drift. This can be conducted using ground-truth data and examining model performance (Section 6) or *without the need for labelling* by other means. These include, for example, statistical tests for out-of-distribution detection [113] and change point detection methods [7]. Drift detection can also be conducted using anomaly detection/novelty detection on incoming data [85].

Some research uses the model parameters themselves to do this and rejects samples for which the model does not have high classification confidence [42]. Care should be taken with such an approach that the measure of confidence is inversely correlated with the novelty of the sample i.e., distance metrics may work well but confidence predictions of models may not [42]; see [114] for a clear use of distance metrics in detecting drift for malware and network intrusion detection. It is also possible to create hybrid algorithms to detect novel samples [76]. Rejecting samples may not be appropriate for all use-cases. It depends on the trade-off cost of a misclassification against no classification, which may be asymmetric for attack detection.

Experimental approaches to altering the training data are not widely adopted or researched.

Readers should be aware that concept-drift mitigation approaches often involve selecting **magic numbers** of some kind in order to fine-tune the drift detection or mitigation. For example, noise reduction requires carefully selecting the amount of noise [38]; drift detection and excluding extreme samples requires choosing threshold parameters. Magic numbers are considered a weakness in a model as they require manual tuning and cannot usually be transferred from one problem to another, but this may be an acceptable trade-off to avoid the cost of labelling data.

Formal methods [110] can be used to verify the expected input-output-mappings of a program. This has also been explored for ML models, but presents a challenge when the input space is large and/or the model is complex (has many parameters), therefore there has been significant research to develop partial formal methods techniques for neural networks in particular. Krichen et al. [54] provide a useful summary of techniques at the time of writing. Partial formal methods may be particularly appropriate for incoming certification to provide some level of assurance of behaviour under particular contexts. See CyBoK Knowledge Area on Formal Methods for more on formal methods [13].

8.6 Bias Mitigation

Bias in machine learning models is a key concern for certification particularly when handling sensitive personal data in case the model is discriminating against certain groups based on features that should not influence the outcome. For compliance with incoming regulation, special attention must be paid to models which process personal data such as UEBA and risk attitude evaluation.

Bias mitigation and explainability practices can help to avoid both spurious correlations and to enable compliance with data governance. Bias can take many forms and is impossible to eliminate entirely. Data scientists may limit learned biases by careful analysis of datasets and selection of model features. Roselli et al. [88] present a clear overview of some of the best practices to limit the introduction of bias into a model.

Explainable AI (XAI) techniques may be implemented to further check for model biases or issues with model robustness. This field of research is still maturing, but many techniques and products already exist to analyse the inputs and outputs of models. Some of these are appropriate for analysis of opaque systems and others require introspection of the model parameters or features.

There are a number of different approaches including: **local explainability** methods (e.g., [87]), which aim to highlight the most relevant features informing a given model output for a *single sample*. Counterfactual explanations are a popular presentation format for local explanations, which help users to identify which features would have to change (and by how much) in order to alter the model prediction. **Global explainability** methods (e.g., [63]) aim to explain which features are most relevant for *all* model outputs.

In using XAI methods, it is critical to consider the audience consuming the explanations and their context - are they experts, non-experts, auditors? Is the explanation required to enable a real-time decision or post-mortem? Some researchers have argued that if explainability is paramount, data scientists should stick to inherently explainable models such as decision-trees, because existing methods cannot explain neural networks behaviour due to their complexity [91]. It is possible for some problems to initially train a neural network and then use transfer learning to create a surrogate model with an inherently explainable algorithm.

9 PROCURING AI SECURITY SOLUTIONS

For ML security products, the trained models are usually components in much-larger systems, which represent valuable intellectual property for the providers and are therefore closed-source. In this section we present some recommendations to test for common pitfalls in opaque systems. This should be caveated that testing of an opaque product including ML will also test any human and rule-based methods included in the wider product. This section focusses *only* on common ML pitfalls, but (i) we cannot be sure that ML is responsible for any failings and (ii) additional tests should always be conducted as in any security product procurement process.

Performance metrics can be misleading. Some best-practice tests for opaque (attack detection) products include:

- Collecting and labelling the *latest* samples that you have collected *in the deployment environment* to ensure (i) the underlying distribution of training data used for the product is not so different that your environment is not well-served (ii) to check generalisation to potentially unseen samples, since attacks with known signatures can easily be filtered out using exclusion lists.
- Obtain an offline copy of the product and do not allow it to update for a period of time, continue to test with new data to test resilience to concept drift over time. Be sure to analyse the range of metrics in which you are interested (Section 6).
- Test with some garbage self-generated inputs to see what happens when truly new data enters the system
- Create some evasive data. Use attack tools, pack or otherwise alter malware that is well-detected by the product, slow down denial-of-service attacks and see if they are still detected. Consider the use of adversarial ML techniques such as universal adversarial patches [56, 19] (see Security and Privacy for AI Knowledge Guide [25] for more on attacking the model)

Understanding that these may not always be practical, ask the vendor to provide:

- the full range of relevant metrics such as false positive and false negative rates, precision (at top-k) and recall for anomaly detection.
- ask for dataset details such as hashes of malware samples, network attacks tested on, anomaly use cases, types of false positives.

Baselining for anomaly detection. If easy to explain and not proprietary, ask how the anomaly threshold is reached. Obtain a trial of the product and generate purposeful anomalies including abnormal but harmless ones to test detection rate and appropriate prioritisation of malicious events. Investigate how the baseline is re-calculated over time and consider a schedule for regularly repeating these tests once deployed.

ML/AI infrastructure often relies on cloud resources to analyse data (e.g. run emulation, make large-scale comparisons), ensure that detection still works when the product runs in offline mode.

Data privacy assurance should be guaranteed for any instances covered by existing governance but sensitive security data may not fall under this legislation and could be incidentally leaked to the vendor and other customers if it is used to train machine learning models. Ask for methodology used to protect customer data. A best practice (but high cost) approach includes creating watermarked samples and attempting to uncover this data from the product.

Robustness, bias and explainability against various failure modes can be tested using the validation techniques described above (Section 9) such as adversarial AI, manipulated attacks, and entering garbage. Ask for global explanations of model behaviour (if not proprietary) and for local explanations of individual sample classifications. Ask for a diagram mapping the technologies in place when the ML enters a failure mode.

10 CONCLUSION

AI is already widely used for attack detection and is increasingly used in other cyber security applications. It is well-suited to analysing large datasets (using machine learning) and automating processes that require frequent updating. Fully autonomous responses to attacks are still considered too risky but it is a highly active research area and this may change in the near future.

Whilst AI is a very broad field that includes explicitly programmed agents, most research and products use machine learning, which infers input-output-mappings from data using ML algorithms to create ML models. It is important to regularly evaluate these models to ensure that privacy and performance are maintained without significant biases and spurious correlations underpinning the technologies.

At the time of writing a number of key pieces of legislation are proposed to govern AI/ML solutions; though the methods for compliance are themselves reliant on immature research fields. Therefore builders of AI/ML systems should be aware that models may be subject to additional requirements in the future, but may try to anticipate these by de-risking models themselves.

REFERENCES

[1] Cyber autonomy gym for experimentation challenge 1. https://github.com/cage-challenge/cage-challenge-1, 2021.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[3] Hussein Abbass, Axel Bender, Svetoslav Gaidow, and Paul Whitbread. Computational red teaming: Past, present and future. *IEEE Computational Intelligence Magazine*, 6(1):30–42, 2011.

[4] Sara Aboukadri, Aafaf Ouaddah, and Abdellatif Mezrioui. Major role of artificial intelligence, machine learning, and deep learning in identity and access management field: Challenges and state of the art. In *Proceedings of the 8th International Conference on Advanced Intelligent Systems and Informatics 2022*, pages 50–64. Springer, 2022.

[5] Baleegh Ahmad, Benjamin Tan, Ramesh Karri, and Hammond Pearce. FLAG: Finding line anomalies (in code) with generative ai. *arXiv preprint arXiv:2306.12643*, 2023.

[6] Orlando Amaral, Muhammad Ilyas Azeem, Sallam Abualhaija, and Lionel C Briand. NLP-based automated compliance checking of data processing agreements against GDPR. *arXiv preprint arXiv:2209.09722*, 2022.

[7] Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

[8] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2017.

[9] Giovanni Apruzzese, Pavel Laskov, and Aliya Tastemirova. SoK: The impact of unlabelled data in cyberthreat detection. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 20–42. IEEE, 2022.

[10] Davide Ariu, Giorgio Giacinto, and Fabio Roli. Machine learning in computer forensics (and the lessons learned from machine learning in computer security). In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 99–104, 2011.

[11] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3971–3988, 2022.

[12] Stefan Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, pages 1–7, 1999.

[13] David Basin. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Formal Methods for Security. University of Bristol, 2021. KA Version 1.0.0.

[14] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for AI. *Communications of the ACM*, 64(7):58–65, 2021.

[15] Ahmed Bensaoud, Nawaf Abudawaood, and Jugal Kalita. Classifying malware images with convolutional neural network models. *International Journal of Network Security*, 22(6):1022–1031, 2020.

[16] Laura M Bishop, Phillip L Morgan, Phoebe M Asquith, George Raywood-Burke, Adam Wedgbury, and Kevin Jones. Examining human individual differences in cyber security and possible implications for human-machine interface design. In *HCI for Cybersecurity, Privacy and Trust: Second International Conference, HCI-CPT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 51–66. Springer, 2020.

[17] Nicolas Botacin. GPThreats-3: Is automatic malware generation a threat? In *Proceedings of the 17th IEEE Workshop on Offensive Technologies (WOOT), co-located with IEEE Security and Privacy*, 2023.

[18] James V Bradley. Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2):144–152, 1978.

[19] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[20] Anna L Buczak and Erhan Guven. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*,

18(2):1153–1176, 2015.

[21] Pete Burnap. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Risk Management & Governance. University of Bristol, 2021. KA Version 1.1.1.

[22] Alvaro Cardenas. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Cyber-Physical Systems Security. University of Bristol, 2021. KA Version 1.0.1.

[23] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[24] Robert Carolina. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Law & Regulation. University of Bristol, 2021. KA Version 1.0.2.

[25] Lorenzo Cavallaro and Emiliano De Cristofaro. Security and privacy of ai knowledge guide. In Awais Rashid, Yulia Cherdantseva, Andrew Martin, and Steve Schneider, editors, *CyBOK Knowledge Guides and Topic Guides*. University of Bristol, 2023. KG Version 1.0.0.

[26] Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. Federated learning for privacy-preserving AI. *Communications of the ACM*, 63(12):33–36, 2020.

[27] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[28] European Commission. 2018 reform of eu data protection rules. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf, 2018.

[29] Roland Croft, Dominic Newlands, Ziyu Chen, and M Ali Babar. An empirical study of rule-based and learning-based approaches for static application security testing. In *Proceedings of the 15th ACM/IEEE international symposium on empirical software engineering and measurement (ESEM)*, pages 1–12, 2021.

[30] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa, et al. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6):e01802, 2019.

[31] Tianhong Dai, Kai Arulkumaran, Tamara Gerbert, Samyakh Tukra, Feryal Behbahani, and Anil Anthony Bharath. Analysing deep reinforcement learning agents trained with domain randomisation. *Neurocomputing*, 493:143–165, 2022.

[32] Hervé Debar. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Security Operations & Incident Management. University of Bristol, 2021. KA Version 1.0.2.

[33] Matthias Eckhart, Andreas Ekelhart, David Allison, Magnus Almgren, Katharina Ceesay-Seitz, Helge Janicke, Simin Nadjm-Tehrani, Awais Rashid, and Mark Yampolskiy. Security-enhancing digital twins: Characteristics, indicators, and future perspectives. *IEEE Security & Privacy*, 2023.

[34] Hoda El Merabet and Abderrahmane Hajraoui. A survey of malware detection techniques based on machine learning. *International Journal of Advanced Computer Science and Applications*, 10(1), 2019.

[35] Simon Yusuf Enoch, Zhibin Huang, Chun Yong Moon, Donghwan Lee, Myung Kil Ahn, and Dong Seong Kim. HARMer: Cyber-attacks automation and evaluation. *IEEE Access*, 8:129397–129414, 2020.

[36] Sascha Fahl. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Web & Mobile Security. University of Bristol, 2021. KA Version 1.0.1.

[37] Sara Cardoso Ferreira. *A machine learning based digital forensics application to detect tampered multimedia files*. PhD thesis, Universidade do Porto (Portugal), 2021.

[38] Tonya Fields, George Hsieh, and Jules Chenou. Mitigating drift in time series data with noise augmentation. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 227–230. IEEE, 2019.

[39] National Institute for Standards and Technology. Framework for improving critical infrastructure cybersecurity. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf, 2018.

[40] Seyed Mohammad Ghaffarian and Hamid Reza Shahriari. Software vulnerability analysis and discovery using machine-learning and data-mining techniques: A survey. *ACM Computing Surveys (CSUR)*, 50(4):1–36, 2017.

[41] Dieter Gollmann. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Authentication, Authorisation & Accountability. University of Bristol, 2021. KA Version 1.0.2.

[42] Jan Philip Göpfert, Barbara Hammer, and Heiko Wersing. Mitigating concept drift via rejection. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, pages 456–467. Springer, 2018.

[43] UK Government. Data protection act 2018. https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted, 2018.

[44] UK Government. Equality act 2010. https://www.legislation.gov.uk/ukpga/2010/15, 2023.

[45] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[46] John Heaps, Ram Krishnan, Yufei Huang, Jianwei Niu, and Ravi Sandhu. Access control policy generation from user stories using machine learning. In *Data and Applications Security and Privacy XXXV: 35th Annual IFIP WG 11.3 Conference, DBSec 2021, Calgary, Canada, July 19–20, 2021, Proceedings 35*, pages 171–188. Springer, 2021.

[47] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[48] Linan Huang and Quanyan Zhu. Adaptive honeypot engagement through reinforcement learning of semi-markov decision processes. In *Decision and Game Theory for Security: 10th International Conference, GameSec 2019, Stockholm, Sweden, October 30–November 1, 2019, Proceedings 10*, pages 196–216. Springer, 2019.

[49] VirusTotal Inc. Virustotal. https://www.virustotal.com/, 2023. Accessed on March 15, 2023.

[50] Jay Jacobs, Sasha Romanosky, Benjamin Edwards, Idris Adjerid, and Michael Roytman. Exploit prediction scoring system (epss). *Digital Threats: Research and Practice*, 2(3):1–17, 2021.

[51] Markus Jakobsson, Elaine Shi, Philippe Golle, Richard Chow, et al. Implicit authentication for mobile devices. In *Proceedings of the 4th USENIX conference on Hot topics in security*, volume 1, pages 25–27. USENIX Association, 2009.

[52] Nathalie Japkowicz. Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning*, 42(1-2):97, 2001.

[53] Igor Kotenko, Elena Fedorchenko, Evgenia Novikova, and Ashish Jha. Cyber attacker profiling for risk analysis based on machine learning. *Sensors*, 23(4):2028, 2023.

[54] Moez Krichen, Alaeddine Mihoub, Mohammed Y Alzahrani, Wilfried Yves Hamilton Adoni, and Tarik Nahhal. Are formal methods applicable to machine learning and artificial intelligence? In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pages 48–53. IEEE, 2022.

[55] Dominik Kus, Eric Wagner, Jan Pennekamp, Konrad Wolsing, Ina Berenice Fink, Markus Dahlmanns, Klaus Wehrle, and Martin Henze. A false sense of security? revisiting the state of machine learning-based industrial intrusion detection. In *Proceedings of the 8th ACM on Cyber-Physical System Security Workshop*, pages 73–84, 2022.

[56] Raphael Labaca-Castro, Luis Muñoz-González, Feargus Pendlebury, Gabi Dreo Rodosek, Fabio Pierazzi, and Lorenzo Cavallaro. Realizable universal adversarial pertur-

bations for malware. *arXiv preprint arXiv:2102.06747*, 2021.

[57] David Last. Forecasting zero-day vulnerabilities. In *Proceedings of the 11th Annual Cyber and Information Security Research Conference*, CISRC '16, New York, NY, USA, 2016. ACM.

[58] Triet HM Le, Huaming Chen, and M Ali Babar. A survey on data-driven software vulnerability assessment and prioritization. *ACM Computing Surveys*, 55(5):1–39, 2022.

[59] Wenke Lee. *The Cyber Security Body of Knowledge v1.0, 2019*, chapter Malware & Attack Technology. University of Bristol, 2019. KA Version 1.0.

[60] Éireann Leverett, Matilda Rhode, and Adam Wedgbury. Vulnerability forecasting: Theory and practice. *Digital Threats: Research and Practice*, 3(4):1–27, 2022.

[61] Liu Liu, Bao-sheng Wang, Bo Yu, and Qiu-xi Zhong. Automatic malware classification and new malware detection using machine learning. *Frontiers of Information Technology & Electronic Engineering*, 18(9):1336–1347, 2017.

[62] Juan Miguel López Velásquez, Sergio Mauricio Martínez Monterrubio, Luis Enrique Sánchez Crespo, and David Garcia Rosado. Systematic review of SIEM technology: SIEM-SC birth. *International Journal of Information Security*, pages 1–21, 2023.

[63] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[64] E. Martin and Tao Xie. Inferring access-control policy properties via machine learning. In *Seventh IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY'06)*, pages 238–242, 2006.

[65] Forrest McKee and David Noever. Chatbots in a honeypot world. *arXiv preprint arXiv:2301.03771*, 2023.

[66] Brad Miller, Alex Kantchelian, Michael Carl Tschantz, Sadia Afroz, Rekha Bachwani, Riyaz Faizullabhoy, Ling Huang, Vaishaal Shankar, Tony Wu, George Yiu, et al. Reviewer integration and performance measurement for malware detection. In *Detection of Intrusions and Malware, and Vulnerability Assessment: 13th International Conference, DIMVA 2016, San Sebastián, Spain, July 7-8, 2016, Proceedings 13*, pages 122–141. Springer, 2016.

[67] Doug Miller, Ron Alford, Andy Applebaum, Henry Foster, Caleb Little, and Blake Strom. Automated adversary emulation: A case for planning and acting with unknowns. Technical report, MITRE CORP, 2018.

[68] Madhav Nair, Rajat Sadhukhan, and Debdeep Mukhopadhyay. Generating secure hardware using ChatGPT resistant to CWEs. *Cryptology ePrint Archive*, 2023.

[69] National Institute of Standards and Technology. AI risk management frameowrk. https://www.nist.gov/itl/ai-risk-management-framework, January 2023. Accessed on March 15, 2023.

[70] Jeffrey Nichols, Kevin Spakes, Cory Watson, and Robert Bridges. Assembling a cyber range to evaluate artificial intelligence/machine learning (AI/ML) security tools. In *ICCWS 2021 16th International Conference on Cyber Warfare and Security*, page 240. Academic Conferences Limited, 2021.

[71] Parliament of Canada. An act to enact the consumer privacy protection act, the personal information and data protection tribunal act and the artificial intelligence and data act and to make consequential and related amendments to other acts. https://www.parl.ca/legisinfo/en/bill/44-1/c-27, 2021.

[72] PRC Ministry of Science and Technology. Ethical norms for new generation artificial intelligence. https://perma.cc/RC4V-Q2FX, 2021.

[73] OpenAI. ChatGPT, 2022. [Software https://chat-gpt.org/].

[74] Ana Lucila Sandoval Orozco, Carlos Quinto Huamán, Daniel Povedano Álvarez, and

Luis Javier García Villalba. A machine learning forensics technique to detect post-processing in digital videos. *Future Generation Computer Systems*, 111:199–212, 2020.

[75] Aissam Outchakoucht, Es-Samaali Hamza, and Jean Philippe Leroy. Dynamic access control policy based on blockchain and machine learning for the internet of things. *International Journal of Advanced Computer Science and Applications*, 8(7), 2017.

[76] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

[77] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[78] European Parliament. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://artificialintelligenceact.eu/the-act/, 2021.

[79] Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[80] Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, Lorenzo Cavallaro, et al. TESSERACT: Eliminating experimental bias in malware classification across space and time. In *Proceedings of the 28th USENIX Security Symposium*, pages 729–746. USENIX Association, 2019.

[81] Sundar Pichai. An important next step on our AI journey, 2023. [Online] https://blog.google/technology/ai/bard-google-ai-search-updates/.

[82] Frank Piessens. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Software Security. University of Bristol, 2021. KA Version 1.0.1.

[83] Romesh Prasad, Matthew K Swanson, and Young Moon. Recovering from cyber-manufacturing attacks by reinforcement learning. In *ASME International Mechanical Engineering Congress and Exposition*, volume 86649. American Society of Mechanical Engineers, 2022.

[84] Abdalbasit Mohammed Qadir and Asaf Varol. The role of machine learning in digital forensics. In *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5. IEEE, 2020.

[85] Hanli Qiao, Boris Novikov, and Jan Olaf Blech. Concept drift analysis by dynamic residual projection for effectively detecting botnet cyber-attacks in IoT scenarios. *IEEE Transactions on Industrial Informatics*, 18(6):3692–3701, 2021.

[86] Matilda Rhode, Lewis Tuson, Pete Burnap, and Kevin Jones. Lab to SOC: robust features for dynamic malware detection. In *2019 49th annual IEEE/IFIP international conference on dependable systems and networks–industry track*, pages 13–16. IEEE, 2019.

[87] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 1135–1144, New York, NY, USA, 2016. ACM.

[88] Drew Roselli, Jeanna Matthews, and Nisha Talagala. Managing bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 539–544, 2019.

[89] Christian Rossow and Sanjay Jha. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Network Security. University of Bristol, 2021. KA Version 2.0.0.

[90] Vassil Roussev. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Forensics. University of Bristol, 2021. KA Version 1.0.1.

[91] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–

215, 2019.

[92] Stuart Russel and Peter Norvig. Artificial intelligence: a modern approach, 2020.

[93] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

[94] T Saranya, S Sridevi, C Deisy, Tran Duc Chung, and MKA Ahamed Khan. Performance analysis of machine learning algorithms in intrusion detection system: A review. *Procedia Computer Science*, 171:1251–1260, 2020.

[95] M. Angela Sasse and Awais Rashid. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Human Factors. University of Bristol, 2021. KA Version 1.0.1.

[96] Burr Settles. Active learning literature survey. https://burrsettles.com/pub/settles.activelearning.pdf, 2010.

[97] Nyle Siddiqui, Laura Pryor, and Rushit Dave. User authentication schemes using machine learning methods—a review. In *Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2021*, pages 703–723. Springer, 2021.

[98] Drew Simshaw. Ethical issues in robo-lawyering: The need for guidance on developing and using artificial intelligence in the practice of law. *Hastings LJ*, 70:173, 2018.

[99] Andrew W Singer. Can AI transform compliance? *Risk Management*, 66(8):4–7, 2019.

[100] Gianluca Stringhini. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Adversarial Behaviours. University of Bristol, 2021. KA Version 1.0.1.

[101] Athor Subroto and Andri Apriyana. Cyber risk prediction through social media big data analytics and statistical machine learning. *Journal of Big Data*, 6(1):50, 2019.

[102] Microsoft Defender Research Team. CyberBattleSim. https://github.com/microsoft/cyberbattlesim, 2021. Created by Christian Seifert, Michael Betser, William Blum, James Bono, Kate Farris, Emily Goren, Justin Grana, Kristian Holsheimer, Brandon Marken, Joshua Neil, Nicole Nichols, Jugal Parikh, Haoran Wei.

[103] Carmela Troncoso. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Privacy & Online Rights. University of Bristol, 2021. KA Version 1.0.2.

[104] Tram Truong-Huu, Prarthana Prathap, Purnima Murali Mohan, and Mohan Gurusamy. Fast and adaptive failure recovery using machine learning in software defined networks. In *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. IEEE, 2019.

[105] Alexey Tsymbal. The problem of concept drift: definitions and related work. *Tech. Rep. Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.

[106] Daniele Ucci, Leonardo Aniello, and Roberto Baldoni. Survey of machine learning techniques for malware analysis. *Computers & Security*, 81:123–147, 2019.

[107] Eva AM van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. ChatGPT: five priorities for research. *Nature*, 614(7947):224–226, 2023.

[108] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

[109] Ingrid Verbauwhede. *The Cyber Security Body of Knowledge v1.1.0, 2021*, chapter Hardware Security. University of Bristol, 2021. KA Version 1.0.1.

[110] Jeannette M Wing. A specifier's introduction to formal methods. *Computer*, 23(9):8–22, 1990.

[111] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[112] Bo Xie, Guowei Shen, Chun Guo, and Yunhe Cui. The named entity recognition of chinese cybersecurity using an active learning strategy. *Wireless Communications and*

Mobile Computing, 2021:1–11, 2021.

[113] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[114] Limin Yang, Wenbo Guo, Qingying Hao, Arridhana Ciptadi, Ali Ahmadzadeh, Xinyu Xing, and Gang Wang. CADE: Detecting and explaining concept drift samples for security applications. In *USENIX security symposium*, pages 2327–2344, 2021.

[115] Shuhan Yuan and Xintao Wu. Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, 104:102221, 2021.

[116] Hanyu Zeng, Zhen Wei Ng, Pengfei Zhou, Xin Lou, David KY Yau, and Marianne Winslett. Detecting cyber attacks in smart grids with massive unlabeled sensing data. In *2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–7. IEEE, 2022.

[117] Hanyu Zeng, Zhen Wei Ng, Pengfei Zhou, Xin Lou, David K.Y. Yau, and Marianne Winslett. Detecting cyber attacks in smart grids with massive unlabeled sensing data. In *2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–7, 2022.

[118] Hanyu Zeng, Zhen Wei Ng, Pengfei Zhou, Xin Lou, David K.Y. Yau, and Marianne Winslett. Detecting cyber attacks in smart grids with massive unlabeled sensing data. In *2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–7, 2022.

[119] Alice Zheng. *Evaluating machine learning models: a beginner's guide to key concepts and pitfalls*. O'Reilly Media, 2015.

[120] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[121] Shuofei Zhu, Jianjun Shi, Limin Yang, Boqin Qin, Ziyi Zhang, Linhai Song, and Gang Wang. Measuring and modeling the label dynamics of online anti-malware engines. In *USENIX Security Symposium*, pages 2361–2378, 2020.

A ACRONYMS

AI Artificial intelligence

AGI Artificial general intelligence

IDS Intrusion detection system

ML Machine learning

NLP Natural language processing

RL Reinforcement learning

SIEM Security information and event management

SOC Security operations centre

UEBA User Entity Behaviour Analytics

XAI Explainable artificial intelligence

B GLOSSARY

- **Action Space** In reinforcement learning, an action space is the set of possible actions from which an automated agent can choose.
- **(Artificial) Neural Networks (ANNs)** are a type of machine learning algorithm loosely inspired by the human brain using networks of 'neurons' to map inputs to outputs.
- **Artificial Intelligence** A broad field studying agents who take inputs from the world, process data and make decisions or act.
- **Baselining** The process of measuring 'normal' activity in a system, often used for anomaly detection.
- **Concept Drift** An evolution of concept accompanied by a change in the underlying data describing that concept.
- **Deep learning** A subfield of machine learning using neural networks which have many internal layers of neurons and connections.
- **Data science** A separate field of research to which machine learning also belongs together with statistics, data science methods are sometimes employed by AI methods.
- **Digital twin** A digitised representation of a system which may exist before the real-world system, updated by live data from the real-world system and/or used as a testbed
- **Differential privacy** A method to both release information from data whilst protecting the privacy of individuals whose data is contained within a database.
- **Deepfake** An artificially generated representation of an individual using static images, audio or video media.
- **Federated learning** A decentralised approach to machine learning whereby data and models are stored in local nodes and used to inform a central AI/ML model without directly sharing the local datasets.
- **Failure mode** A way in which a system might cease to function within desirable operating parameters.
- **Formal Methods** A way to mathematically specify the expected behaviours of a program or system which can be used for systematic verification. See [110].
- **Homomorphic encryption** Enables useful manipulation and analysis on ciphertext such that results can be extracted from encrypted data.
- **Input-output-mappings** Generic rules, decisions of the AI/ML agents/models.
- **Machine learning** A subfield of AI reliant on data in order to learn patterns and produce outputs such as labels, probabilities and new data samples itself.
- **Performance metrics** Used to describe the degree to which a model achieves a specific goal; often as a fraction or percentage over many examples.
- **Symbolic reasoning** One subfield of AI that embeds human expert knowledge (facts and logic) and then processes new facts and logic to give outputs.