On the Need for Collaborative Intelligence in Cybersecurity

Trevor Martin [1]

[1] *Machine Intelligence Unit, Engineering Maths, University of Bristol, Bristol, BS8 1UB, UK*

Abstract

The success of artificial intelligence (and particularly data-driven machine learning) in classifying and making predictions from large bodies of data has led to an expectation that autonomous AI systems can be deployed in cybersecurity applications. In this position paper we outline some of the problems facing machine learning in cybersecurity and argue for a collaborative approach where humans contribute insight and understanding, whilst machines are used to gather, filter and process data into a convenient and understandable form. In turn this requires a convenient representation for exchanging information between machine and human, and we argue that graded concepts are suitable, allowing summarisation at multiple levels of discernibility (granularity).

Keywords

Cybersecurity, collaborative intelligence, explainability

1. Introduction

Recent developments have pushed artificial intelligence to the forefront of many applications in areas as diverse as product and service recommendation, autonomous and partially autonomous vehicles, healthcare, finance, voice-driven interfaces / interactive systems, telecom routing, and fraud detection, amongst many others.

This development has been made possible by vast increases in our ability to store, transmit and process data, underpinned by a similarly large leap in the number and capabilities of devices able to gather data, and the general population's acceptance (or ignorance) of the data gathering processes. It is worth pointing out that popular use of the term "artificial intelligence" nowadays frequently refers to the sub-area of statistical or data-driven machine learning that is able to capture patterns in the data and use them to make decisions (often predictions or classifications) in new cases.

AI has been identified as having the potential to change both sides of the cyber-security landscape, both attack and defence. The rapid increase in data-driven AI classifiers and decision-makers is a double-edged sword - it can make more powerful and more adaptive defences, but can also increase the sophistication of an attack.

However it is important to note that AI (and in particular, data-driven machine learning) is not a "silver bullet" solution to the problems of detecting anomalous activity in cybersecurity. The strength of machine-learning tools is finding activity that is similar to something previously seen, without the need to precisely describe that activity; cybersecurity problems generally involve an adversary who is trying to subvert a system in a novel way, using methods that are new or not previously detected. Deep learning systems are vulnerable to so-called adversarial attacks, where a carefully crafted input can lead to completely incorrect classifications. Great care is needed in ensuring that training data is clean and represents the whole space.

It is difficult to completely define an anomaly in the context of cyber-security - it is an event (or sequence of events) that is not necessarily rare, but that does not conform to expected behaviour and possibly constitutes a threat. Of course, this begs the question of what constitutes expected behaviour

AI-CyberSec 2021: Workshop on Artificial Intelligence and Cyber Security, December 14, 2021, Cambridge, UK
EMAIL: trevor.martin@bristol.ac.uk

- frequently, a baseline of data is taken as normal behaviour, and subsequent data is examined for anomalies. If the initial dataset can be guaranteed clean (and representative), this approach is valid, but if it is possible the initial data could contain "anomalous" instances then it risks classifying malicious events as normal. The boundaries of normal behaviour are rarely clear-cut, and may change with time - particularly in cyber-security applications when adversaries often aim to avoid raising suspicion, and can adapt to updated defences. The key to a good cyberattack is undetectability - this distinguishes anomaly detection in cybersecurity data from most applications of anomaly detection, since the presence of an adversary may mean that baseline data can rarely be guaranteed clean.

We therefore consider that fully autonomous cyberdefence is neither likely nor desirable. Whilst many routine tasks can be automated, the overall security of a network or system is an evolving problem that is not amenable to standard machine learning tasks. Indeed, the paucity of available cyberattack data actively works against data-driven machine learning. The possibility of inadvertently releasing personal details (even in anonymised records) is a strong motivation to withhold data. In addition, there is an imperative to stop a cyberattack when it is detected, rather than allowing it to continue so that representative data can be gathered. There are datasets showing simulated attacks but these are somewhat artificial and not necessarily representative of real attacks.

Our view is that collaborative intelligence - a symbiotic combination of humans and machines - is a better approach than autonomous AI to monitor and proactively control the behaviour of a complex system such as a computer network. Even when autonomous decisions are necessary to cope with the pace of events, retrospective analysis demands human involvement. The idea of collaborative intelligence can be traced back to multi-agent systems where computational tasks were distributed amongst processors - either in a heterogeneous manner where each component had a specific “expertise” or in a homogeneous fashion, where a problem was divided into similar, but smaller, sub-problems, processed separately. By allowing "agents" to be human or machine, we arrive at a collaborative intelligence approach. This relies on simple, effective and efficient communication of information between the “processing components” i.e. computers and human analysts. A key feature in enhancing human understanding of large scale data is the notion of summarisation. By combining many values (or objects) into a few entities, concise summaries enable analysts to gain insight into bulk patterns and focus on the mechanisms underlying relations in the data, leading to greater understanding of current data and prediction of future data.

In the remainder of this position paper, we outline some of the key background material and suggest that use of graded concepts is a promising approach to information sharing between humans and machines.

2. The use of AI techniques for anomaly detection.

Anomaly detection predates AI, and is often studied as a sub-field of statistics. Frequently, anomalies are equated with outliers and a variety of statistical techniques exist to detect outliers (and/or noise) in data, both static and dynamic (time-series).

It is common to label as outliers any data points that deviate significantly from common statistical properties such as mean, median, etc., or points that fall into particular quantiles. Two commonly quoted definitions of an outlier are:

“… an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism” [1]

and

“… an observation or subset of observations that appears to be inconsistent with the rest of the set of data”. [2]

The difficulty of detecting anomalies / outliers is often increased in streaming applications [3], where, typically, data can only be viewed once, and data volume/velocity may mean that detailed analysis is not possible. Additional problems arise from so-called "drift and shift" where properties such as mean (within a specified window) can vary systematically over time, e.g. daily temperatures might exhibit significant seasonal variation, complicating the detection of anomalous values.

Anomalies can be sub-divided into

- single anomalies (one data point stands out from the rest),
- contextual anomalies (conditional) where a data point is not unusual in itself, but is not expected in the surrounding set of data - particularly in a time series,
- collective anomalies where a set of data points appear to be unusual.

The use of anomaly detection for cyber-security applications is obvious, and goes back to at least the 1980s [4], where statistical profiles were proposed to identify anomalous behaviour, and rule-based actions were taken on detection of anomalies.

2.1. What is an Anomaly

It is worth pointing out the (possibly obvious) fact that an anomaly is not just an example that differs from others. In a sense, all distinct examples are unique, and as more features of the data are considered, there are more ways to distinguish any given instance. To take a trivial example, consider the list "*hen, sheep, cow, dog*" and identify the anomaly (odd one out). There are plausible reasons for identifying any member of the list as anomalous - hen is the only bird, dog is the only carnivore, cow is the only word with its letters in alphabetic order, sheep is the only 5 letter word, etc. This may seem a contrived example, but there is an important point here - given a set of examples in a sufficiently rich feature space, it is normally possible to make a case for any example being out of the ordinary and hence anomalous.

There are several good surveys of the anomaly detection problem, both from the general perspective [5, 6] and from the more specialised viewpoint of cyber security [7]. For the statistician, an anomaly is an outlier but not all outliers are anomalies. A common statistical starting point (e.g. [8]) is to assume that data is produced by an underlying random process and to distinguish outliers (data produced by the random process, with low probability) from anomalies (data produced by another process). Of course, this is not applicable if the data can't be modelled as a random process, or if the process cannot be known with sufficient accuracy to distinguish between data arising from the model with low probability and data that is not from the model.

2.2. Artificial Intelligence and Anomaly Detection

Anomaly detection is a natural application for data-driven machine learning. Two popular approaches are

- one-class statistical learning - if we view anomalies as rare events, it makes sense to attempt to learn the patterns of normal behaviour and treat anything outside this class as an anomaly- see [9] for an example. We would argue that the one-class approach is not always applicable for cyber-security applications - it essentially splits data into normal and abnormal, but attempts to learn the classifier only from examples of normal data. Caution is required, since a typical machine learning program needs a large number of representatives of all classes that are to be detected, and when there is only one "normal" class, the "abnormal" class may consist of several disjoint subclasses (including some with no available examples).
- techniques based on deep-learning, particularly use of reconstruction error to indicate anomalies. Essentially this relies on building an auto-encoder using a dataset the defines normal behaviour, and then running the auto-encoder with new data. For data close to the original set, the auto-encoder will produce very similar data as a good approximation to the input; if the predicted input was a long way from the actual input then actual input is likely to be anomalous. A broader overview of "deep anomaly detection" is presented in [10] which includes a section focused on anomalies in intrusion detection systems and a list of datasets. Problems related to computational complexity and noisy / unbalanced training data are acknowledged, as is the difficulty of retraining to cope with changes in the data stream. Explanation is not considered in the survey. See also [11] for a view of explainable deep learning in process monitoring

Evangelou and Adams [12] present a framework for cybersecurity anomaly detection based on NetFlow records, where an anomaly is defined as an increased number of connection to other hosts, relative to a baseline of "normal" behaviour. Scalability of the approach is not clear - the study is based on data from the Imperial College network which is quoted as having in the region of 40000 devices connected to it each day; the study rests on analysing 55 randomly selected devices.

Duan et al [13] apply reinforcement learning to detect unexpected events in log sequences from distributed database and distributed processing applications. These have some similarity to cyber-security logs, although the approach relies on the notion of an event sequence being defined and detectable. This is generally the case in database transactions but may be less well-defined in cyber-security logs.

Provision of representative labelled data is a major issue for many approaches in cyber-security - as mentioned above, a competent adversary may be able to pollute normal data, and once an anomaly (intrusion, etc.) is detected, good practice aims to block the vulnerabilities identified as soon as possible, so that there is limited opportunity to gather examples of the anomalous incident [14]. The lack of data is not just a problem for machine learning applications that rely on adequate sources of data for training; it also causes problems for validation and comparison of different approaches to anomaly detection. As discussed in [15], confidentiality is the main reason - real-world data can reveal highly sensitive information, even when it is seemingly sanitised.

Sommer and Paxson [15] also identify problems in using machine learning for anomaly detection (particularly network intrusions):

> *"despite extensive academic research one finds a striking gap in terms of actual deployments of such systems: compared with other intrusion detection approaches, machine learning is rarely employed in operational "real world" settings"*

Amongst the difficulties identified by these authors are the high cost of errors, lack of adequate training and testing data, and a semantic gap between classification results and operational interpretation (that is to say, once an event is identified as anomalous, there is no indication whether it is an isolated incident arising from user error, etc., or is part of an attack and requires action). The authors also highlight the possibility that in some circumstances it might be quicker and easier to look for a simple, non-machine learning approach.

2.3. Explainability in Anomaly Detection

The ubiquity of data-driven intelligent systems has led to a demand for so-called "explainability" [16]. The vast majority of automated decisions are made on the basis of large datasets and complex internal processing that can be incomprehensible even to domain experts. In many cases we need to understand why a particular output is given, rather than just focus on the accuracy of the output. The DARPA initiative [17] is widely credited with sparking much of the research in the area of statistical machine learning, although the broader area of explainability in AI has a longer history, going back at least to the second AI boom of the 1980's.

There are a number of good surveys and overviews in explainable AI, including [18]; [19]; [20]).

Explainability is a key feature of anomaly detection - whether AI based or not - due to the need to understand and (possibly) react to anomalies. As with explainability in AI, it is possible that a human will immediately see why an anomaly has been labelled; it is also possible that the reason(s) will not be immediately apparent and in order to distinguish between anomalies and false alarms, reasons for the classification are required. An explanation can also assist in building confidence in the anomaly detection method, and allow flaws to be detected where false positives are generated. Over-generation of false positives is a particular problem for anomaly detectors, and can rapidly lead to lack of trust in a deployed system, given that substantial work is often required to investigate an alarm.

The approach used in [8] finds outliers and anomalies; the authors use an "isolation forest" and "active anomaly discovery" (AAD) to improve identification of anomalies by incorporating user feedback. The isolation forest approach starts with a data set (numerical) and repeatedly selects a feature at random, then splits on values until there is only one point at each leaf. Regular data requires many more splits than outliers, hence anomalies are usually found at a shallow tree node. The process

is repeated for (e.g.) 100 trees, and the AAD system assigns anomaly scores and asks for user judgment on the highest ranked cases before re-calculating weights to downgrade outliers and upgrade anomalies according to the expert categorisation.

The work is extended in [21] to detect anomalies in event logs, by classifying each 30-minute period (over 2 weeks) as anomalous or not, depending on the count of various events; by presenting the top 20 anomalous time intervals to an analyst, the scoring of anomalous events was refined in two stages and the number of false positives reduced (the authors state that this is due to feedback although it is possible that better results were due to looking at different sets within the data)

Saad et al [22] consider the role of machine learning in malware detection, suggesting there is an over-optimistic view of AI as a solution to malware attacks. They point out that

> *" malware attacks in the wild continue to grow and manage to bypass malware detection systems powered by machine learning techniques. This is because it is difficult to operate and deploy machine learning for malware detection in a production environment or the performance in a production environment is disturbing (e.g., high false positives rate). In fact, there is a significant difference (a detection gap) between the accuracy of malware detection techniques in the literature and their accuracy in a production environment"*

The authors suggest use of if-then rules based on the input features used by a black-box classifier to approximate the malicious/benign decisions; such a representation is assumed to be adequate for communication with analysts.

[23] indirectly address the issue of explainability by proposing a system that automatically classifies anomalies identified by an IDS into a taxonomy, thereby assisting analysts in determining the seriousness of the anomaly and possible actions that can be taken.

Explainability is also a key factor identified by [24], who point out that analysts must be able to identify true positives in a large volume of noisy alerts from IDS and other sources, and rapidly decide whether the events constitute an attack, the nature of the attack, likely evolution of the attack, etc. Their proposed system aims to automate the initial filtering by learning a set of state machines from observation of analyst actions, so that knowledge is elicited and can be examined / transferred / reused.

The idea of using a framework based on human-understandable concepts is also introduced by [25], in the context of malware detection. In common with others, they point to the gap between laboratory studies and real-world deployment of machine-learning systems in cyber security, and use a concept learning system based on description logic to develop tools to distinguish malware from benign binary files. This has the advantage that human expert knowledge can also be encoded and easily combined with the machine-generated knowledge. This can be interpreted as an anomaly-detection task, and provides a pointer to a more knowledge-based and understandable representation that could be useful in wider cyber-security anomaly detection.

Visualisation is also an important aspect of understanding how / why anomalies occur. The system described in [26] relies on external detection systems to find intrusions and gives insight into the communication patterns of the legitimate and suspicious traffic by means of high quality graphics - in this case, summarising NetFlow records on a large scale.

3. Routes to Explainability

Three primary approaches can be identified to explainable systems:

1. "Transparent-by-design" refers to systems that use a naturally understandable representation such as decision trees or rule-based classifiers. This rests on an assumption that symbolic representations are relatively easy to understand. The flaws in this assumption will be readily apparent to anyone who has tried to understand a convoluted SQL query, rule set, decision tree (or even a prolog program); however, for shallow trees and rule sets based on "natural" splits and groupings in the data, there is a good possibility that a human would be able to follow and/or highlight flaws in the reasoning process. For example, it should be relatively straightforward to understand why a medical diagnosis system produces a diagnosis of "common cold" by citing runny nose, sore throat, etc. as reasons for the diagnosis; one that explains its conclusion by

producing a list of several hundred numerical variables and their allowed ranges would require considerably more insight and thought to make sense of. Case-based reasoning is another example of this category - it rests on an assumption that if a human can see why a decision has been reached in some typical cases, then they should be able to interpolate or extrapolate from those examples to a "similar" case.

2. Systems can be designed with explainability as a core component. This can be achieved by restricting the complexity of a machine learning model, or by building into the design features that support explanation. There are few (if any) software packages that follow this approach. Given that software libraries supporting black box models are relatively available and easy to use, writing bespoke software (or rewriting existing code) is generally a high cost route to explainability.

3. The most common approach is to use a black-box method in the first instance to develop a good classifier / predictor, and then adopt a "model-agnostic approach" [27] to build an explanation component as an add-on to the system. A post-hoc explanation system fits one or more transparent-by-design components to each region of the input-output space [28]. For example, a set of decision trees could be built to replicate the input -output behaviour in restricted cases, with each tree handling a small part of the overall input space. Within each tree, it should be possible to see the key features and values that lead to a particular conclusion; hence it can be claimed that the system is interpretable. A similar interpretability case can be made for small rule sets, where the chosen features and attribute values indicate the important features, or a linear model where the important features can be identified by the weights.

We focus mainly on the third approach, implicitly including aspects of the first approach since "transparent-by-design" representations are the normal route to post-hoc explanations.

Case-based reasoning can be a useful representation for post-hoc explanation, since the specific cases chosen as exemplars should be well understood and, provided that the similarity metric is good, the parallel with a new case will make an obvious explanation for the decision. Keane [29] uses case-based reasoning as an add-on to a neural net classifier and provides a review of the coupling between case-based reasoning and neural nets (including deep learning systems), highlighting their origins in coupled neural net/case-based systems developed in the late 1990s.

It is, of course, possible to regard the black box model as a mapping from inputs to outputs, and attempt to understand (for an individual case) which of the inputs are most important in determining the output in that particular case. A popular approach, SHAP [30] uses Shapley values from game theory to assess the contribution of each input variable and hence to indicate which variables or combination of variables lead to each decision.

The use of post-hoc interpretable models is popular in XAI literature as it is relatively easy to optimise the post hoc model against a metric (such as tree size) and to argue that the metric acts a proxy for interpretability. Although this makes research easier (and more reproducible) it is potentially misleading [31] as there is only intuitive support for a link between interpretability and tree compactness / explanation length / etc.

In general, the systems mentioned above assume the user has some expertise and understanding of the problem domain. Formalising this understanding by means of an ontology can feed into the explanation process - for example, in [32], decision trees are used to reproduce neural net results on two benchmark datasets and the authors demonstrate that (in these cases at least) decision trees refined with ontological knowledge are syntactically less complex and are better for human understanding than the unrefined versions. In this work, the ontology is used specifically to estimate the information content of a concept, and more general concepts are preferred as nodes in the trees.

We note also that the post-hoc, individual case based approach can be useful for local explainability, i.e. understanding the input-output relation on a case-by-case basis, but is less suitable for global explainability, i.e. understanding the classification behaviour over the whole space The global approach is arguably more important for a human to understand when developing knowledge of the system behaviour (a cognitive model), and hence a degree of trust in the system's decisions. Lakkaraju et al [33] argue that from this viewpoint, the idea of global model interpretability is more important than local interpretability and use so-called 2-level decision sets, in which the second level is a simple tree / rule set and the first level decides which second level to use.

In order for a human to develop a global understanding of a model, it is necessary to accurately predict outputs for any given input. As the internal workings of an algorithm become more complex, understanding the reasons for a particular decision becomes increasingly difficult and verifying that the decision is "correct" becomes impossible. At some point, it becomes easier to experiment with different input data than to understand the internal workings of a system.

Finally, a key point that is often neglected concerns the interface for explanation. In post-hoc interpretability, the knowledge representation is fixed at design time, either as a decision tree, or a rule set, linear model etc. However, the HCI community has identified richer and more expressive forms of interaction such as natural language (possibly restricted to some kind of formal query language), visualisation, and more specialised approaches such as symbolic mathematical models. There appears to be a great deal more "explanatory power" in an interactive process, where the user can request further detail or an alternative perspective on some aspects of the explanation. This is closer to a human-to-human mode of explanation but requires considerably more effort than today's simpler approaches. It represents an information transfer beyond the simple decision output, such as the input parameters that could be changed in order to change the decision, cause-and-effect links within the model, and general trust / confidence in the model output.

3.1. Explainability and Trust

The ultimate aim of explainability in a human-computer collaboration is to build trust in the machine's decisions and recommendations. It is clear that human teams function better when individuals trust each other's judgments and decisions, so we can expect a similar performance advantage when humans trust the automated agent(s) in a team. However, the notion of trust in an AI systems is difficult to define, quantify and measure - possibly even more difficult than explainability and interpretability - so we will leave this as an open question. It is interesting to speculate that trust could be expressed as a function of an AI system's accuracy and explainability (on the grounds that we trust a system if it gives correct answers and we understand why it has given those answers). In this case, it might be possible to estimate the degree to which a system is explainable from its accuracy and the degree of trust a user has in the system.

3.2. Evaluation

Assuming a satisfactory definition of explainability can be agreed, the obvious question for system evaluation and comparison is how to measure it (in cybersecurity or on the more general AI context).

There are many proposals, ranging from simple syntactic measures such as tree depth (for decision trees), through to complex experiments based on human factors and task efficiency. The former group is generally easy to quantify although it should be noted there is no direct relation between tree depth / rule length and understandability. There is an intuitive link between the two, in that a simple tree or rule set is likely to be more easily understood than a complex set; hence a proxy measure of tree / rule complexity can provide an indication but not a guarantee of interpretability. The second approach - evaluating interpretability by measuring overall system effectiveness - is much more difficult to set up and to repeat. It relies on a notion that if decisions are of better quality with the aid of the automated component, then the system must be understandable.

Expanding this view, Doshi-Velez and Kim [34] suggest that interpretability should be evaluated via one of three general approaches :

- evaluating the system in its intended application, comparing outputs (decisions) with and without the explanation system to judge whether it improves overall performance or not. For example, a medical decision-making system could be evaluated by comparing patient outcomes
- evaluating the system on simplified applications that represent the real application in some way - for example, asking humans to judge the best explanation from several candidates
- using a proxy evaluation (such as tree depth), assuming that there is evidence linking the proxy to an application-based evaluation. In this way, it should be possible to avoid optimising the system without regard to real world performance

In a study more focused on black box classifiers, Backhaus and Seiffert [35] proposed 3 criteria to compare interpretability :

- ability to indicate important input features
- ability to provide typical data points representing a class
- existence of model parameters that indicate the location of decision boundaries

The issue of interpretability can also be viewed in the light of moving beyond test-set accuracy as the primary measure of a classifier system, to considering its use in practical situations and developing metrics for usability. In particular (see [34]) there is little need for interpretability if there is no serious consequence of a wrong decision or if the system accurately reproduces the judgement of a human. For example, recognising numerals is a process that can be automated but the results are relatively easy for a human to validate.

In some senses the issue of judging interpretability of a system has many parallels to the Turing test and (as with the Turing test), a convenient approach is to turn the problem round. In the case of the Turing test, the question of how to define intelligence is replaced by a practical test that if a human judge cannot tell the difference between a computer and human response to arbitrary questioning, the computer must be classed as intelligent. In the case of interpretability, the equivalent argument states that if a system can satisfactorily explain its reasoning to a human then it is interpretable. This, of course, merely shifts the problem of evaluating a system's interpretability to evaluating the quality of an explanation.

The observation that explainability is a property of system and user, not just of the system, is fundamental to the analysis and survey presented in [36]. It is neatly summarised in the quote:

"explanations are social — they are a transfer of knowledge"

Nevertheless, the majority of studies that attempt to measure explainability fall back on looking at a simple proxy measure (such as tree complexity). In the area of neural nets coupled to case-based explainers surveyed by Keane [29], fewer than 5% carried out user-trials, a situation referred to as "the embarrassment of user testing" and highlighted as one of three main areas for future research.

3.3. History

Explainability is not a new problem for AI - the expert systems of the 1980's faced similar demands, even though they were largely based on symbolic representations and built by acquiring and codifying knowledge from human experts rather than from data. It seems that many current research projects linked to explainability are ignoring the experience and lessons available from this earlier work. Moore [37] is a frequently quoted review that surveys many of its contemporary systems and highlights the shortcomings of the simple "print out a trace of rules used" approach to explaining why a particular piece of data was requested or how a conclusion was reached.

As discussed in [38], three core issues governed the effectiveness of an expert system in real-world use - knowledge acquisition, knowledge representation, and the communication interface. Data driven systems are not concerned with the first aspect, since knowledge (or its equivalent) is extracted from data rather than from a domain expert; however, the remaining two components are highly relevant in explainability. Kidd observed that the knowledge representation must be able to capture the range and power of an expert's knowledge, and must be compatible at a cognitive level with the expert's view of the problem domain. In particular, simple rules were often found to be inadequate to express constraints, heuristics, causal links and interactions, and procedural knowledge. Furthermore, behaviour of rule-based systems is often implicitly dependent on internal factors such as the order in which rules are tried.

Regarding the communication interface, both Kidd and other workers (see for example, the discussion of mycin in [37]) noted the need for a flexible and two way "dialogue" so that exploration of "why" and "why not" questions could take place, and a full understanding of the reasoning process could be obtained.

This is not to say the explainability problem was solved at the time - typically, explanation was implemented in a simple manner, by presenting the rules and data used in deriving a conclusion. This approach that relied on the user understanding the representation and content of the knowledge base.

A number of studies at the time highlighted the need for better interfaces and exchange between human and computer. For example, [39] reported that doctors using a medical expert system gave the highest rating to the "ability to explain diagnostic and treatment decisions", third highest to the ability to "display an understanding of their own medical knowledge", and rated "never make an incorrect diagnosis" as 14th out of 15 desirable properties.

The lesson for today's explainability researchers is that so-called understandable representations such as rules, decision trees, case-based reasoners, etc. may not be sufficiently expressive to convey the real reasoning underlying a data-driven decision-making process. Attention must also be paid to " cognitive aspects of the user interface, including dialogue control, explanation facilities, user models, natural language processing"[40]. The observation that

> *" Explainers must have alternative strategies for producing responses so that they may provide elaboration or clarification when users are not satisfied with the first explanation given. Furthermore, the system must be able to interpret follow-up questions about misunderstood explanations or requests for elaboration in the context of the dialogue that has already occurred, and not as independent questions."*

is just as relevant to today's AI systems as it was to the expert systems of the past.

4. Beyond Simple Representations

Again drawing on lessons from the past, Michalski [41], in an early paper on inductive learning (which can be regarded as a fore-runner of data-driven learning), commented that

> *"the results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single 'chunks' of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion"*

This should be a guiding principle for explainable AI - the ability to express concepts in terms of simpler concepts, and to combine those concepts into higher level concepts as well as generating causal explanations. A concept encompasses a class of entities and constraints, properties, relations etc. that distinguish the class of entities from other entities in the domain. This indicates the need for some kind of formal representation of a domain - not necessarily a full-blown ontology but at least a taxonomy of entities, and agreement on the scope and meaning of commonly used terms. Whilst not explicitly recommending use of an ontology, Weihs [42]introduced the ideas of mental fit and data fit, and highlighted the need to match concepts used in rules to those terms that make sense to the user.

A number of other studies highlight the requirement to define terms and concepts as a pre-requisite for exchanging knowledge between human and computer components of a system. The research reported in [43] starts from the common view that representations such as decision trees are inherently understandable, and that a decision tree (possibly using different splits in different areas of the input space) is a suitable "explanation" for a decision. They use decision trees to reproduce neural net results on two benchmark datasets and demonstrate that (in these cases at least) decision trees refined with ontological knowledge are less (syntactically) complex and, in some limited user-based evaluations, are better for human understanding than the unrefined versions. They use an ontology to define a generalisation / specialisation of concepts and prefer trees that have more general concepts. Based on trials they find these decision trees are more understandable than the unrefined versions. The understandability is assessed in three ways, by presenting examples to users (for manual classification using the tree), by presenting generic statements, what-ifs (what would you need to change in order to get a different outcome), and by pairwise tree comparison.

Riveiro and Thill [44] build on the idea that classification systems should use the same concepts as a user in explanation, suggesting additionally that " explanations should align with end user expectations " - in other words, when a user expects a particular output but the system produces a different output, the explanation should address why the expected output wasn't produced as well as why the actual output was produced. Importantly, they found their approach helped users to gain an understanding of the AI system by building an accurate mental model, providing additional evidence for the view that a shared understanding is a necessary component for effective explainable AI. This should be a global model, covering all cases, rather than a local model based on a few examples - for

example, Chromik et al. [31] performed a set of experiments that illustrated how focusing on a few specific cases and simplified "local models" could be counter-productive, in leading users to a false or misleading idea of how the decisions are reached.

Chari et al [45] provide a comprehensive historical overview of explainability in knowledge-based systems from the early 70s through the present day. They highlight the need for "Provenance-aware, personalized, and context-aware explanations", suggesting that the future lies in multi-actor systems which do not rely on the thought processes of a single individual (or the outputs of a single classifier system), but instead focus on the exchange and transformation of knowledge between multiple actors. The ability of the computer-based components to provide explanations, respond to queries and generally reference relevant information is a vital ingredient, and can only be implemented by a rich framework for knowledge representation within which humans and computer-based systems can interact Provenance of explanations is an important, and often neglected, issue . Information about the domain knowledge and sources of the data used in arriving at a decision is a key component in gaining user trust. In [45], provenance is proposed as a key component of explainability in knowledge systems, building on work in the semantic web using ontologies and knowledge graphs to structure and reason about explanations. The paper discusses taxonomies to classify explainable AI systems (using the IBM research framework) This includes the idea of an explanation taxonomy, which enables designers of ML models to include features that will assist post-hoc model interpretation. Note that other researchers such as Gilpin et al.[19] also propose taxonomies to categorise explanation capabilities, and Sokol and Flach [46] go further in providing multiple taxonomies for classification and comparison of different approaches under five distinct headings or "dimensions" (functional requirements, operational requirements, usability requirements, safety requirements, validation requirements). However, Chari et al make a stronger case than most for the use of knowledge structures such as ontologies as part of the explanation process, rather than as a means of categorising or comparing the explanation process.

4.1. Graded Knowledge Representation for Collaborative Intelligence

Machine processing is generally centered on well-defined entities and relations, ranging from the flat table structures of database systems through graph-based representations and up to ontological approaches involving formal logics. On the other hand, human language and communication is based on a degree of vagueness and ambiguity that leads to an efficient transmission of information between humans without the need for precise definition of every term used. Even quantities that can be measured precisely (height of a person or building, volume of a sound, amount of rainfall, colour of an object, etc.) are usually described in non-precise terms such as tall, loud, quite heavy, dark green, etc. More abstract properties such as beautiful landscape, delicious food, pleasant weather, clear documentation, corporate social responsibility, are essentially ill-defined, whether they are based on a holistic assessment or reduced to a combination of lower-level, measurable quantities.

Zadeh's initial formulation of fuzzy sets [47] was inspired primarily by the flexibility of definitions in natural language. He argued that most natural language terms (concepts) admit "graded membership", in that it is possible to compare two objects and to say whether or not one belongs more strongly to the concept. Clearly in the case of an elementary quantity such as height, we are generally able to say that person-1 satisfies the concept "tall" better than person-2 (or that they satisfy the concept equally well). Such gradation can be confirmed by measurement, if necessary. However, it is also valid to speak of graded membership in the case of more complex concepts such as those listed above. We can generally rank the membership of different objects in the set representing the concept extension - in other words, the concept extension can be modelled as a fuzzy set. The interval [0, 1] is a convenient range for the membership function. It maps naturally to a scale where definite membership can be represented by 1, non-membership by 0, with intermediate values used to reflect the lack of a precise border between the two extremes. However, the fundamental idea is that membership is ordered, not the precise membership on a scale.

In addition to the use of flexible terms, human reasoning is characterised by an ability to switch between different levels of granularity when dealing with a problem

More recent work [48, 49] has combined the notions of formal concept analysis with graded tolerance relations. A concept is a set of objects, which cannot be distinguished on the basis of the describing attributes (intension). Given a partition of attribute values, equivalence classes naturally divide the objects into non-overlapping sets, each of which contains objects that are indiscernible on the basis of the attribute. For graded partitions, we obtain a nested sequence of lattices. We form a concept lattice using standard techniques - however, the lattice varies according to the membership grade used to create the partition.

4.2. Simple Example

A specific application area is in the analysis of cyber attacks where we use an ontology such as https://capec.mitre.org/ to categorise events into hierarchical classes. The category labels are fuzzy (in the sense that events may belong more or less strongly to a category), and the approach described in [49] allows us to extend existing software to the case of fuzzy membership in categories, without needing to modify the underlying software. Fig 1 shows a fragment of the concept lattice used.

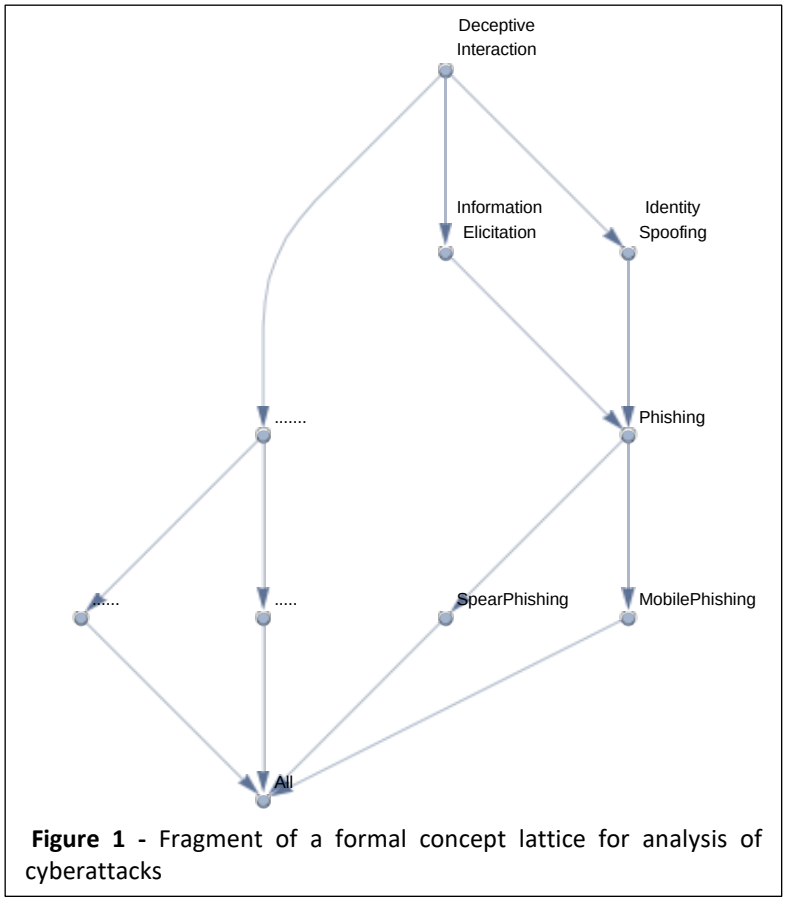

Figure 1 - Fragment of a formal concept lattice for analysis of cyberattacks

5. Summary

The use of collaborative AI in cybersecurity requires seamless exchange of knowledge between agents, whether human or machine. Explainability is not a property of an intelligent system. It is a function of (at least) 3 elements - the task(s), the computer-based components in the process and the human participants. We should not spend time on philosophical discussions of explainability, but focus on whether or not the requisite tasks can be completed more effectively by the collaborative (computer + human) system when "explainability" is included in the collaborative interface. Graded formal concepts are an enabler for the exchange of information between humans and machines in a collaborative intelligent system, allowing us to model many of the soft definitions used in natural language descriptions of events.

6. References

[1] D. M. Hawkins, *Identification of outliers*: Springer, 1980.

[2] V. Barnett and T. Lewis, "Outliers in statistical data," *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*, 1984.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection for Discrete Sequences: A Survey," *IEEE Trans on Knowledge and Data Engineering*, vol. 24, pp. 823-839, 2012.

[4] D. E. Denning, "An Intrusion-Detection Model," *IEEE Transactions on Software Engineering*, vol. SE-13, pp. 222-232, 1987.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, 2009.

[6] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier Detection for Temporal Data: A Survey," *IEEE Trans on Knowledge and Data Engineering*, vol. 26, pp. 2250-2267, 2014.

[7] N. Hoque, M. H. Bhuyan, R. C. Baishya, D. K. Bhattacharyya, and J. K. Kalita, "Network attacks: Taxonomy, tools and systems," *Journal of Network and Computer Applications*, vol. 40, pp. 307-324, 2014.

[8] S. Das, W.-K. Wong, A. Fern, T. G. Dietterich, and M. A. Siddiqui, "Incorporating Feedback into Tree-based Anomaly Detection," *CoRR*, vol. abs/1708.09441, / 2017.

[9] N. Moustafa, *et al.*, "DAD: A Distributed Anomaly Detection system using ensemble one-class statistical learning in edge networks," *Future Generation Computer Systems*, vol. 118, pp. 240-251, 2021/05/01/ 2021.

[10] R. Chalapathy and S. Chawla. (2019, January 01, 2019). Deep Learning for Anomaly Detection: A Survey. arXiv:1901.03407. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190103407C

[11] K. Amarasinghe, K. Kenney, and M. Manic, "Toward Explainable Deep Neural Network Based Anomaly Detection," in *2018 11th International Conference on Human System Interaction (HSI)*, 2018, pp. 311-317.

[12] M. Evangelou and N. M. Adams, "An anomaly detection framework for cyber-security data," *Computers & Security*, vol. 97, p. 101941, 2020/10/01/ 2020.

[13] X. Duan, S. Ying, W. Yuan, H. Cheng, and X. Yin, "QLLog: A log anomaly detection method based on Q-learning algorithm," *Inf. Processing & Mgt*, vol. 58, p. 102540, 2021.

[14] O. Yavanoglu and M. Aydos, "A review on cyber security datasets for machine learning algorithms," in *IEEE International Conference on Big Data (Big Data)*, 2017, pp. 2186-2193.

[15] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *IEEE Symp. on Security and Privacy*, 2010, pp. 305-316.

[16] D. Castelvecchi, "Can we open the black box of AI?," *Nature*, vol. 538, pp. 20-23, 2016.

[17] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, vol. 40, pp. 44-58, 06/24 2019.

[18] A. Barredo Arrieta, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Infor. Fusion*, vol. 58, pp. 82-115, 2020.

[19] L. Gilpin, *et al.*, "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning," in *IEEE Intl.Conf. Data Sci.and Adv. Analytics*, 2018, arXiv:1806.00069.

[20] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, vol. 8, p. 832, 2019.

[21] M. A. Siddiqui, *et al.*, "Detecting Cyber Attacks Using Anomaly Detection with Explanations and Expert Feedback," in *IEEE Intl Conf. (ICASSP)*, 2019, pp. 2872-2876.

[22] S. Saad, W. Briguglio, and H. Elmiligi, The Curious Case of Machine Learning In Malware Detection, *Int. Conf. on Information Systems Security and Privacy* Prague, pp. 528-535. 2019.

[23] D. Bolzoni, S. Etalle, and P. H. Hartel, "Panacea: Automating Attack Classification for Anomaly-Based Network Intrusion Detection Systems," Berlin, Heidelberg, 2009, pp. 1-20.

[24] C. Zhong, J. Yen, P. Liu, and R. F. Erbacher, "Automate Cybersecurity Data Triage by Leveraging Human Analysts' Cognitive Process," in *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity/HPSC/IDS)*, 2016, pp. 357-363.

[25] P. Svec, S. Balogh, and M. Homola, "Experimental Evaluation of Description Logic Concept Learning Algorithms for Static Malware Detection," 2021.

[26] F. Fischer, F. Mansmann, D. A. Keim, S. Pietzko, and M. Waldvogel, "Large-Scale Network Monitoring for Visual Analysis of Attacks," Berlin, Heidelberg, 2008, pp. 111-118.

[27] M. Ribeiro, S. Singh, and C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, 2016.

[28] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS*, 2017.

[29] M. T. Keane and E. M. Kenny, "How Case-Based Reasoning Explains Neural Networks: A Theoretical Analysis of XAI Using Post-Hoc Explanation-by-Example from a Survey of ANN-CBR Twin-Systems," in *Case-Based Reasoning Research and Development*, 2019, pp. 155-171.

[30] L. Antwarg, R. Mindlin Miller, B. Shapira, and L. Rokach. (2019, March 01, 2019). Explaining Anomalies Detected by Autoencoders Using SHAP. arXiv:1903.02407. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190302407A

[31] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz, "I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI," presented at the 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, 2021.

[32] R. Confalonieri and T. R. Besold, "TREPAN Reloaded: A Knowledge-Driven Approach to Explaining Black-Box Models," in *ECAI*, 2020.

[33] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable & Explorable Approximations of Black Box Models," *CoRR*, vol. abs/1707.01154, / 2017.

[34] F. Doshi-Velez and B. Kim. (2017, Towards A Rigorous Science of Interpretable Machine Learning. *arxiv eprint 1702.08608*.

[35] A. Backhaus and U. Seiffert, "Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size," *Neurocomputing*, vol. 131, pp. 15–22, 05/01 2014.

[36] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2019/02/01/ 2019.

[37] J. Moore and W. Swartout, "Explanation in expert systems: A survey," 01/01 1989.

[38] A. L. Kidd and M. B. Cooper, "Man-machine interface issues in the construction and use of an expert system," *International Journal of Man-Machine Studies*, vol. 22, pp. 91-102, 1985.

[39] R. L. Teach and E. H. Shortliffe, "An analysis of physician attitudes regarding computer-based clinical consultation systems," *Computers and Biomedical Research*, vol. 14, pp. 542-558, 1981.

[40] D. C. Berry and D. E. Broadbent, "Expert systems and the man–machine interface," *Expert Systems*, vol. 3, pp. 228-231, 1986/10/01 1986.

[41] R. S. Michalski, "A theory and methodology of inductive learning," *Artificial Intelligence*, vol. 20, pp. 111-161, 1983/02/01/ 1983.

[42] C. Weihs and U. M. Sondhauss, "Combining Mental Fit and Data Fit for Classification Rule Selection," in *Exploratory Data Analysis in Empirical Research*, Berlin, 2003, pp. 188-203.

[43] R. Confalonieri, T. Weyde, T. R. Besold, and F. Moscoso del Prado Martín, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artificial Intelligence*, vol. 296, p. 103471, 2021/07/01/ 2021.

[44] M. Riveiro and S. Thill, "“That's (not) the output I expected!” On the role of end user expectations in creating explanations of AI systems," *Artificial Intelligence*, vol. 298, p. 103507, 2021.

[45] S. Chari, D. M. Gruen, O. Seneviratne, and D. L. McGuinness, "Foundations of explainable knowledge-enabled systems," *arXiv preprint arXiv:2003.07520*, 2020.

[46] K. Sokol and P. Flach, "Explainability fact sheets: a framework for systematic assessment of explainable approaches," presented at the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 2020.

[47] L. A. Zadeh, "Fuzzy Sets," *Information and Control*, vol. 8, pp. 338-353, 1965.

[48] T. P. Martin and B. Azvine, "Graded associations in situation awareness," in *2017 Joint 17th International Fuzzy Systems Association (IFSA-SCIS)*, 2017, pp. 1-6.

[49] T. P. Martin and B. Azvine, "Graded Concepts for Collaborative Intelligence," in *Proc. IEEE Intl. Conference on Systems, Man, and Cybernetics, SMC 2018*, 2019, pp. 2589-2594.