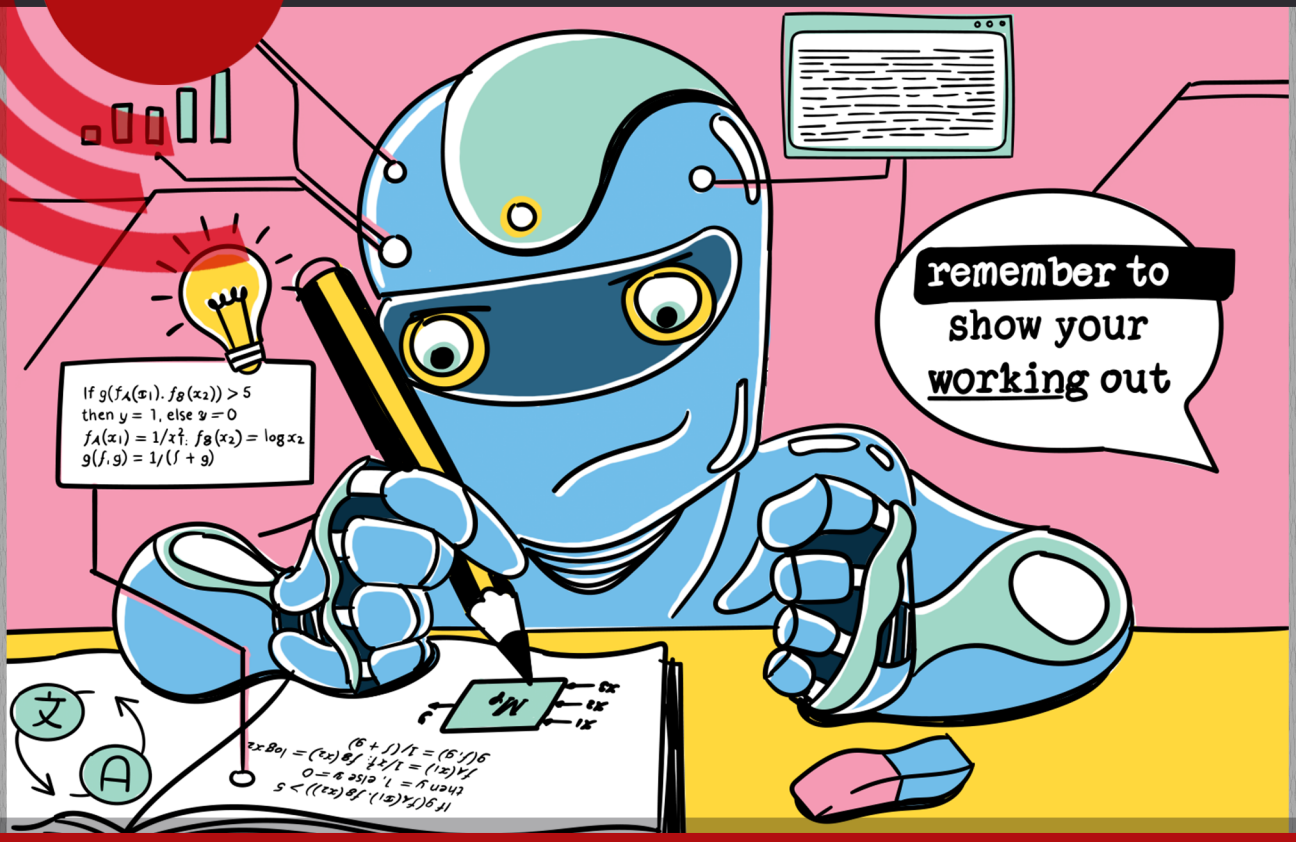
CREST Centre for Research and Evidence on Security Threats



Understanding The Problem Of Explanation When Using Artificial Intelligence In Intelligence Analysis

FULL REPORT

JUNE 2021

Chris Baber
Emily McCormick
Ian Apperley

JUNE 2021

Understanding The Problem Of Explanation When Using Artificial Intelligence In Intelligence Analysis

FULL REPORT

Chris Baber, Emily McCormick, Ian Apperley
University of Birmingham

This research was funded by the Centre for Research and Evidence on Security Threats – an independent Centre commissioned by the Economic and Social Research Council (ESRC Award: ES/N009614/1) and which is funded in part by the UK security and intelligence agencies and Home Office.

www.crestresearch.ac.uk

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

This report explores the role of explanation in human engagement with artificial intelligence / machine learning (AI / ML).

For AI / ML to augment human intelligence (in terms of extending a human's cognitive capabilities through the provision of sophisticated analysis on massive data sets), there needs to be sufficient common ground in the way humans and AI / ML communicate.

In this report, we assume that interactions between humans and AI / ML occur in a system in which cooperation between humans and AI / ML is one interaction among many, e.g. humans cooperate with other humans; humans programme the AI / ML; humans could be involved in selecting and preparing the data that the algorithms use; the AI / ML could interact with other algorithms etc.

Not only is it important that humans and AI / ML establish common ground, but also that humans who communicate with each other using AI / ML share this common ground.

From this perspective, the term 'explanation' is the process by which common ground between interactions is established and maintained.

We have developed a framework to highlight this concept, and this is instantiated to show how different types of explanation can occur, each of which requires different means of support.

Primarily, an explanation involves an agreement on the features (in data sets or a situation) which the 'explainer' and 'explainee' pay attention to and why these features are relevant.

We propose three levels of relevance:

- 'Cluster' – In which a group of features typically occur together
- 'Belief' – which defines a reason as to why such a cluster will occur
- 'Policy' – which justifies the belief and relates this to action.

Agreement (on features and relevance) depends on the knowledge and experience of the explainer and 'explainee', and much of the process of the explanation involves ensuring alignment between parties in terms of knowledge and experience.

We relate the concept of explanation developed here to concepts such as intelligibility and transparency in the AI / ML literature and provide guidelines that can inform decisions on the development, deployment, and use of AI / ML in operational settings.

From the framework of explanation developed in this report, we propose the following guidelines:

1. **Explanations should include relevant causes**
   Explanations should relate to beliefs in the relationship between features of a situation and the causes that can directly affect the event being explained (probability) or can explain most of the event (explanatory power); are plausible (construct validity); and if the cause was instigated by a person, deliberative.
2. **Explanations should include relevant features**
   Explanations should relate to the key features of the situation and the goals of the explainer and explainee.
3. **Explanations should be framed to suit the audience**
   Explainers should fit the explanation to suit the explainee's understanding of the topic and what it is they wish to gain from the explanation (their

mental model and goals).

4. **Explanations should be interactive**
   Explainers should involve explainees in the explanation.
5. **Explanations should be (where necessary) actionable**
   Explainees should be given information that can be used to perform and / or improve future actions and behaviours.

2. PROJECT OVERVIEW

The primary approach taken in this project is the development of a framework that can be used to predict how different types of explanation are developed and used. This is presented in *Section 5*.

In support of this approach, we conducted literature reviews of the concept of explanation in the human science (*Appendix A*) and in machine learning (ML) (*Section 4*). The literature reviews were complemented by workshops at the National Computer Security Centre in London. Two workshops were run, one in late November and one in early December 2019. Subsequent workshops were cancelled as the result of changes in working and travelling due to Covid-19. At each workshop, participants were drawn from a variety of organisations involved in security, i.e. including the Ministry of Defence, police forces, financial technology, and computer security.

Each workshop involved up to eight participants and lasted from around 10am to 4pm (with an hour break for lunch). We did not, for obvious reasons, collect any demographic data from participants (apart from the first names they provided) and any information presented in this report has been sanitised for 'Official' security level.

The workshops began with a brief introduction to the project in terms of the project objectives (although we did provide much detail in order not to lead participants or the discussion). Following this, participants were asked to work in pairs or groups of three to discuss their experience of an activity in which data analytics (which might include ML) were deployed. This was captured in the form of Post-it notes (one for each task or step in the activity), which were laid out on the table to create simple process flow models of the activity

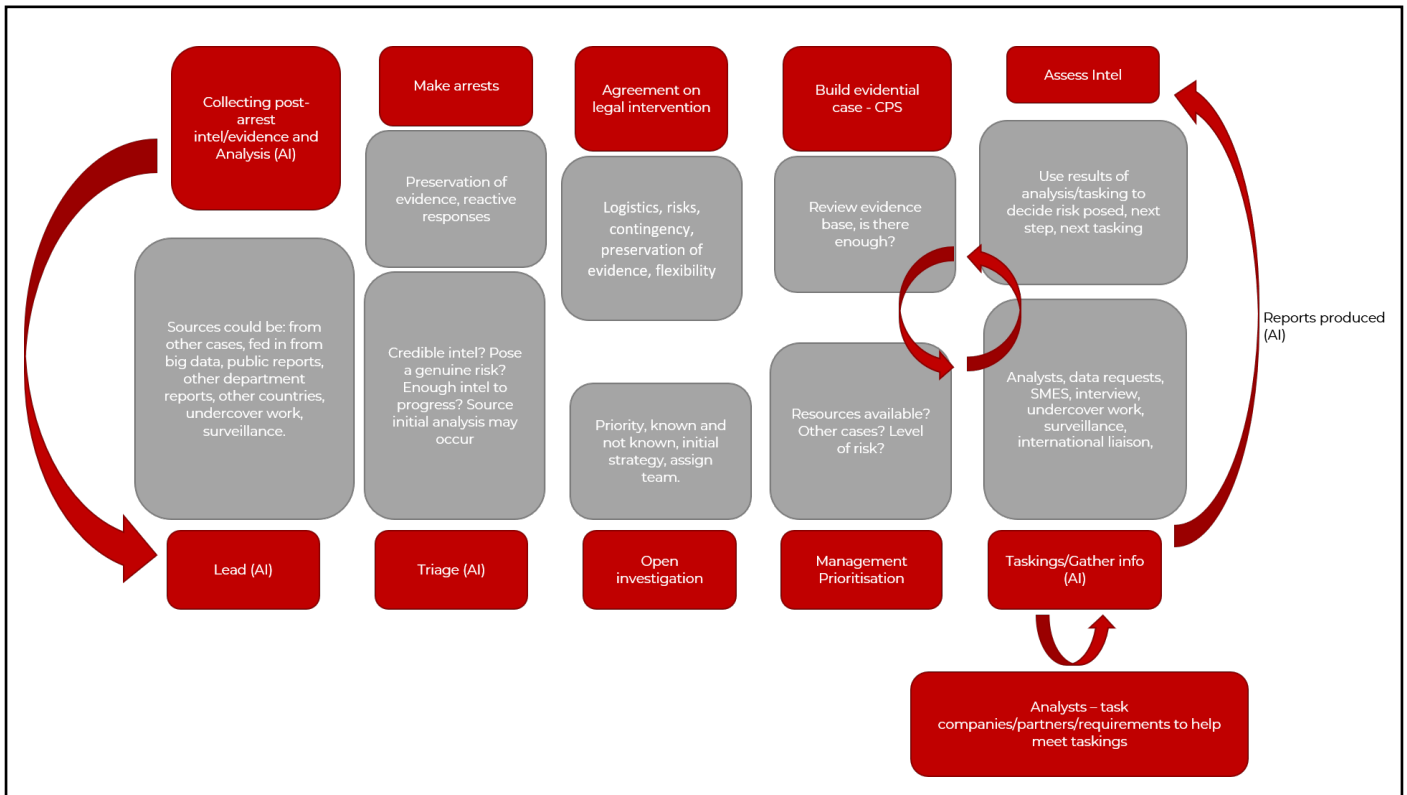


Figure 1. Process model derived from workshop activity to describe intelligence analysis for disrupting trafficking

(*Figure 1*). Versions of these are presented later in the report.

Broadly, the activities focused on identifying and preventing; countering or otherwise reducing threats; or increasing situation awareness or strategic advantage over adversaries. From the process flow models, participants were asked to indicate the stakeholders (individuals, organisations, computers) that would either perform tasks or would be affected by tasks in these activities. This ranged from the operatives (or sensors) collecting intelligence, through to analysts and to people involved in compiling, disseminating, and reading reports arising from these activities.

For example, a report might, via journalists, be read some time after the activity by members of the public, or a summary report with high security classification might be read by a Senior Investigating Officer as part of a daily briefing, or a detailed analysis (perhaps with an even higher security rating) might be read by fellow analysts as part of the ongoing analysis.

The examples highlighted the need to balance information security with the information needs of the stakeholders (and to ensure that undue inferences could not be drawn from whatever information was made available to stakeholders). These points are developed further in the report.

Participants were also asked to consider "what makes a 'good' explanation?" and "what are the benefits of ML?" They wrote single words or short phrases on Post-it notes, and these were grouped into 'affinity diagrams', i.e. placing related concepts together (*Figure 2*). Participants were invited to either add to these or to move the concepts themselves until there was consensus among the group that the resulting collections were organised appropriately. While the workshops were kept as open and qualitative as possible, several participants shared experiences of the activities that were discussed and perspectives on positive or negative deployment of ML (or other data analytics).

From the workshops, explanation often relies on the assumption that there is a relationship between an outcome (effect) and cause(s). This relationship could be defined in terms of relevance (to the situation, to

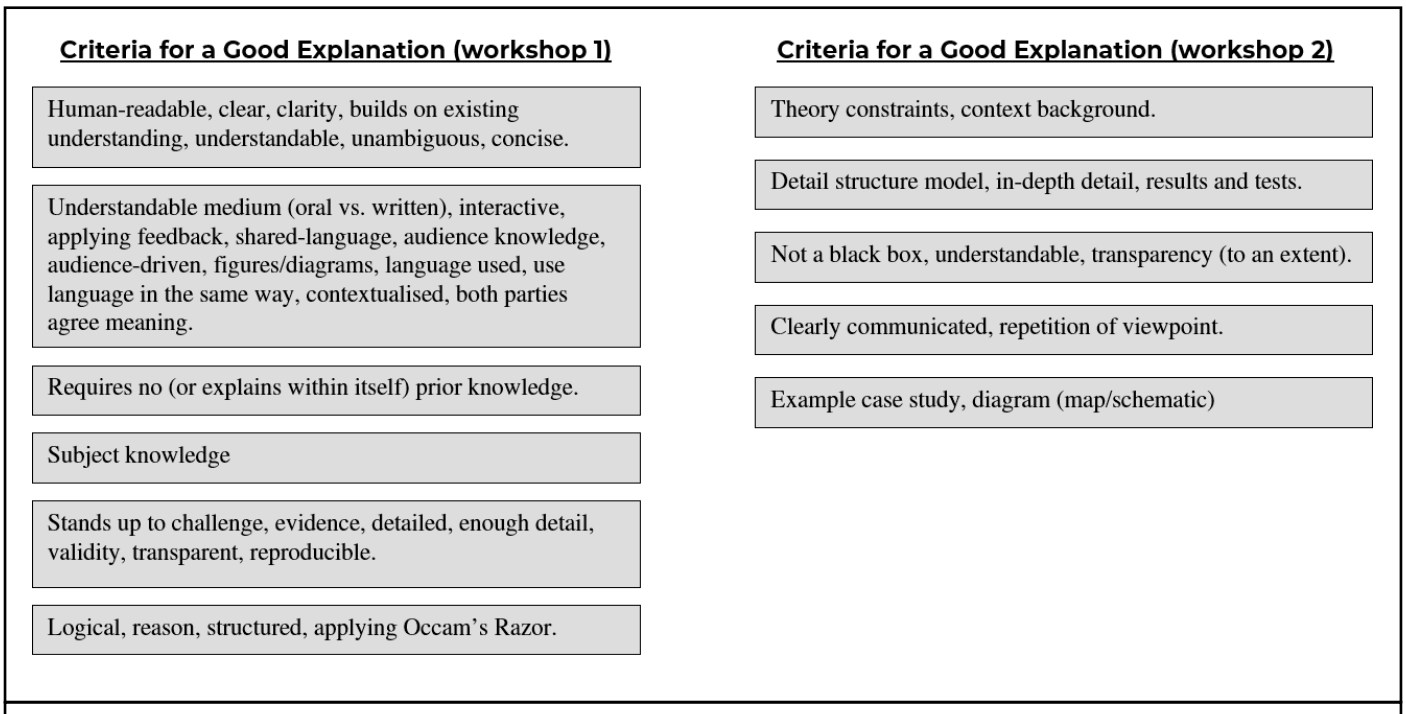


Figure 2. What makes a good explanation as decided by attendees of the two workshops. Participants were also asked to group the words together based on whether they felt they fell under the same category.

the explainer, or to the explainee). In this account, explanation can be divided into discrete but connected stages: In a specific situation, the explainer will explore known (or available) features, which combines sense-making of the situation and the predicted consequences of the explanation for the explainee in terms of the assumed mental model, goals, and abilities of the explainee.

Following this, features are selected in terms of their relevance and an explanation product is defined. The quality of an explanation can be evaluated (by explainer or explainee) and this could occur in a social context (in which explainer and explainee could discuss or otherwise negotiate the explanation until it was acceptable). In some circumstances, the explainer might anticipate this negotiation by modifying the content or structure of the explanation to better suit their model of the knowledge held by the explainee.

Overall, the focus group emphasised the importance of a transparent, understandable, and relevant explanation, together with the importance of using a shared language during an explanation. For example, only using terms and concepts that both the explainer and explainee are familiar with. This was largely parallel to what the literature stated, however, there were some key similarities and differences that we noticed.

One major theme identified from analysing the workflows was that good explanation should be an interactive process, actively involving both the explainer and explainee(s). For instance, in *Figure 1*, the stages involved in disrupting trafficking are cyclical (as noted in the sense-making and intelligence analysis in *Section A.6*). This also implies that, at least to some extent, there is interaction between different parties involved in the process. This idea accords with Clark's (2015) notion of common ground and Sperber and Wilson's relevance theory (2002) (*Section A.2*).

However, while many workshop attendees recognised barriers that prevented the bidirectional flow of information, it was apparent that explanation is still necessary and can still occur. This supported the notion

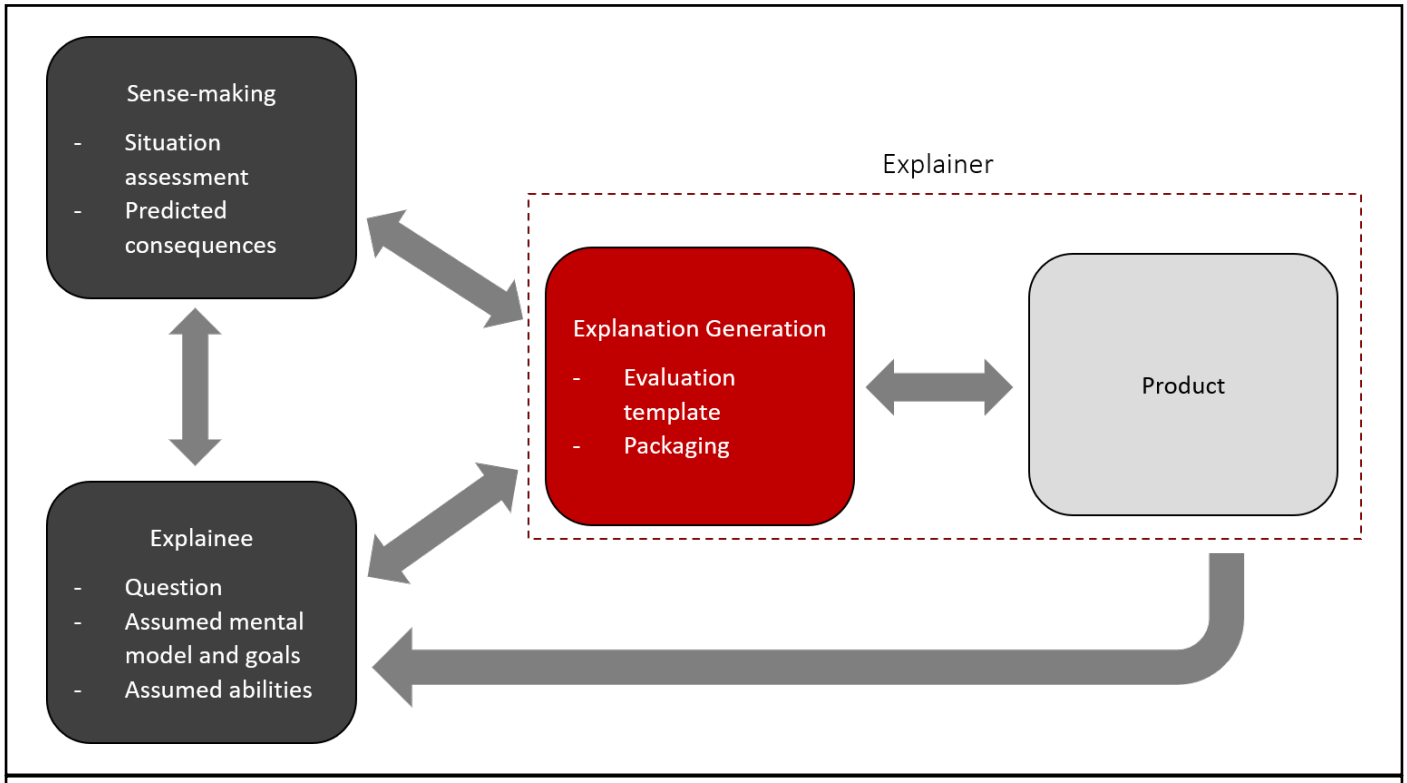


Figure 3. Initial framework of how explanations are generated in humans

that explainers create some model of the explainee and use this as a reference point when generating an explanation. Therefore, subsequent improvements to the framework considered how it would work and under different social situations and how the role of the explainee would change considering these.

Another aspect regarding explanation that was apparent in both the literature and the workshop was that there are different 'levels' of explanation. These different levels not only depended on the question asked, but the context the explanation occurred in. So, the environmental factors surrounding the explanation impacts the explanation produced.

According to Hoffman et al. (2018), explanations can be separated into local and global categories. Local explanations are used to rectify a flaw in the explainee's understanding, whereas a global explanation is used to broaden an explainee's understanding of a topic (Klein et al., 2019; Hoffman et al., 2018).

However, after hearing feedback from the focus groups, we felt that the levels of explanation ran deeper than this. Since the workshop attendees emphasised in discussions that an explanation must be actionable to be useful, we felt that there should be a separate level of explanation. In this level, the explainer must consider the actions of the explainee when generating the explanation. Before elaborating these points, we turn to the notions of explanation applied in ML.

We have used the information provided to complement our literature review (*Appendix A*) and to create a framework for explanation (*Section 6*).

3. THE NEED FOR EXPLANATION IN AI SYSTEMS

According to a 2017 report from the AI Committee of the British Parliament:

> *"The development of intelligible AI systems is a fundamental necessity if AI is to become an integral and trusted tool in our society… Whether this takes the form of technical transparency, explainability, or indeed both, will depend on the context and the stakes involved, but in most cases we believe explainability will be a more useful approach for the citizen and the consumer.… We believe it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take…. In cases such as deep neural networks, where it is not yet possible to generate thorough explanations for the decisions that are made, this may mean delaying their deployment for particular uses until alternative solutions are found".*[1]

AI systems should be capable of explaining any decisions that they make to people who will be affected by these decisions. Article 22 of the EU General Data Protection Regulation (GDPR) and the European Commission's Ethics Guidelines for Trustworthy AI both emphasise that people have a right to an explanation from automated decisions that might affect them. The Fairness, Accountability, and Transparency in Machine Learning (FATML) campaign for accountable algorithms[2] proposes that explainability should be able to present details, in non-technical language, on how an algorithm has reached a specific decision to any stakeholder (i.e. any individual who might be affected by that decision) if they ask for it. Additionally, the FATML guidelines note that there should be clear indication of who is responsible for the decision (particularly for any failures), that the algorithm should allow auditability of the decision process, should indicate the accuracy of data and process (i.e. in terms of sources of error and uncertainty), and should ensure that any decisions are fair (i.e. not biased against any demographics).

At root, these regulations and guidelines share the overarching goal of ensuring that any decision made using AI or ML should be explained to the humans who will act upon or be affected by the decision. One approach to the challenge of explanation is to focus on the algorithm itself:

> *"Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand."*
>
> *(Arrieta et al., 2020).*

While early forms of AI were sufficiently simple to have their rules open to inspection, recent developments in ML and AI have led to systems that are 'black boxes' from which it can be difficult to extract explanations that humans can understand (particularly if this is to occur with minimal knowledge of the inner workings of these algorithms). Broadly, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves

1 "AI in the UK: Ready, Willing and Able?," report, UK Parliament (House of Lords) Artificial Intelligence Committee, 16 April 2017; https:// publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002 .htm.
2 https:///www.faml.org.resoureces/ principles-for-accountable-algorithms

with experience E." [Mitchell, 1997, p.2]. From this, the learning arises from the training experience through which data are added to improve a model's performance. This points to two factors that define the model: the first is the underlying statistical structure that defines the theoretical scope of a solution; the second is the algorithm in which this statistical structure is implemented. Let's assume that the task, T, involves assigning an image to one category or another.

The statistical structure assumes that the image can be decomposed into features, and that some (but not necessarily all) of these features will be strongly related to one of the categories. How the selection of features or definition of relationship is implemented will depend on the algorithm. Even before progressing with the report, think about how this process (of image decomposition followed by correlation of features to categories) might differ from visual perception performed by humans when recognising an image as, say, a cat rather than a dog.

One of the differences is that humans are likely to use their prior knowledge and experience of categories in ways that ML algorithms do not (i.e. very few ML algorithms make use of the sort of common sense that humans apply). Another difference lies in the granularity at which humans or ML algorithms define a feature (i.e. humans will not decompose an image into its constituent pixels). While these differences ought to be obvious to the reader, they illustrate some of the stumbling blocks in the path towards explainable algorithms.

Put simply, the definition of a feature; the process by which features are extracted and analysed; the underlying statistical structure that relates features to categories; and the processes by which these relationships are implemented are entirely different.

From this perspective, should an explanation focus on the definition of features, the underlying statistical structure, and the implementation of the algorithm?

3.1 EXPLANATION AND RELATED TERMS

To understand what humans might require of explanation in AI, this project develops a framework for explanation inspired by research in the human and social sciences. In broad terms, explanation of an algorithm can be considered in line with a host of concepts in which a decision-maker provides an account of the rationale for that decision. Some examples of these concepts are given in *Table 1*.

Intelligibility	Human can understand the underlying algorithm
Understandability	See above
Comprehensibility	Human can understand the knowledge used by the algorithm
Interpretability	Human can understand the meaning of an algorithm's output
Transparency	See Intelligibility. Also, algorithm can be explored
Decomposability	Each element (step) of an algorithm has intelligibility
Simulatability	Operation of the algorithm can be imagined by human
Responsibility	Human ultimately responsible for action arising from analysis

Table 1. Concepts related to AI / ML explanation

There is some overlap of definitions in *Table 1.* For example, intelligibility, transparency, decomposability, and simulatability each refer to the human's ability to understand how an algorithm is applied to data. In this respect, 'explainability' is related to the person's knowledge of, and competence in, computer science / mathematics (which could limit the explainability of the algorithm for people who do not possess such knowledge). For example, *Figure 1* shows the steps required by a human to make sense of a ML algorithm (M) in a given state, ϕ, given a set of variables $(x_1 \ldots x_n)$, and an output y.

In simulatability (*Figure 4a*), the human can understand the formulae that underpin the algorithm and perform calculations to produce identical results. In decomposibility (*Figure 4b*), the human can trace the effect of the different variables on the algorithm, perhaps by calculating output for different values of these variables. In transparency (*Figure 4c*), the human can interpret the algorithm's output in terms of rules the define relationships between variables.

The relationship between explainability and a person's knowledge of ML / AI, suggests that there is a possibility that a lack of technical knowledge of the algorithm can result in misconceptions [AI Literacy [25]] or the use of a black box algorithm can produce misunderstandings [AI Literacy [55]]. Having said that, there is evidence that simply visualising the workings of the algorithm is, even for experienced users of ML / AI, insufficient to overcome problems in detecting, correcting, and responding to errors in the algorithm [Interpreting interpret[8]]. Indeed, such visualisation might encourage superficial evaluation (because the picture might feel plausible) and, as long as the output aligns with the analyst's intuitions, the result might be accepted without question [Interpreting interpret…]. Even when the intuitions work against the output of the algorithm, and the analyst is suspicious, a visualisation might be insufficient to enable a detailed argument to be formulated.

One (possibly tongue-in-cheek) definition of explainability we encountered during conversations relating to this project was that an explanation (of a deep neural network) was simply the vector that describes the parameter weightings once the algorithm had settled on a solution.

Such a perspective feels contrary to the aspiration for algorithms to provide explanation to any stakeholder. Having said that, contemporary neural networks (DNN, RNN, CNN etc.) do not easily lend themselves to either explanation of process or rationale for output. In this sense, even aside from the algorithm, it can be difficult to provide comprehensibility because the knowledge used by the algorithm might not be surfaced in a format that a human is able to perceive.

For example, a neural network trained to distinguish the difference between a panda and a gibbon (*Figure 5*) might focus on elements in the images at the pixel level, which are not perceptible to human viewers, and might group these into elements which are not meaningful to a human-understandable concept of either of these entities. Given this pixel-level analysis, spoofing or otherwise interfering with the ability of image recognition algorithms is now a popular activity in AI communities (and forms the basis of Adversarial Neural Net research in which pairs of AI systems are pitted against each other to disrupt recognition – primarily to learn robustness against such attacks).

If explanation cannot be based solely on the workings of the algorithms, then what other perspective could we take? The output of an algorithm could be considered

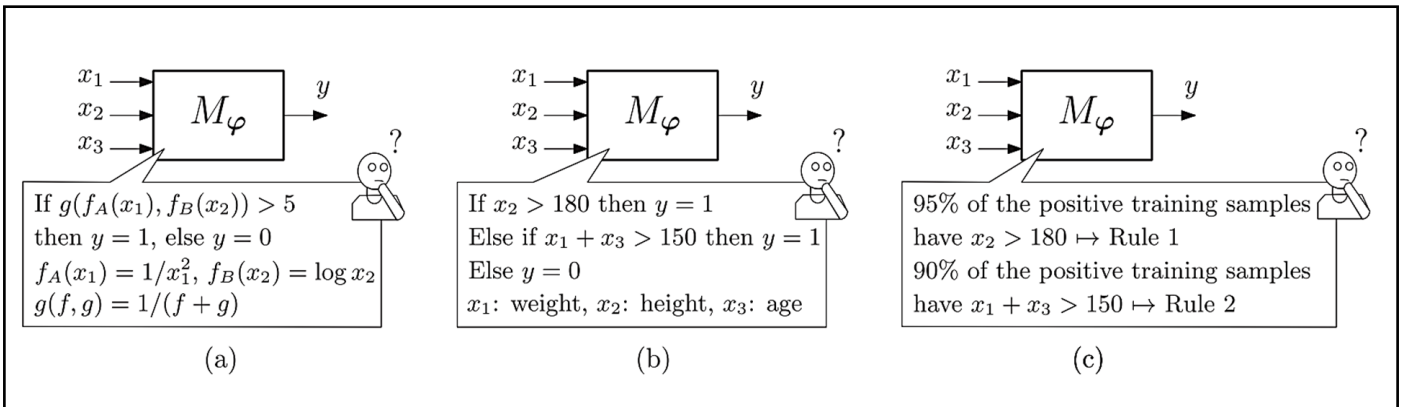


Figure 4. Examples of an algorithm's simulatability (a), decomposibility (b), and transparency (c) for a human observer [Arietta et al., 2020]

in terms of its ability to produce a result that the human stakeholder could evaluate. As *Table 2* implies, such evaluation could involve the algorithm being run on a set of data that were different from those on which it was trained. In this manner, the performance of the algorithm could be checked. An assumption here is that, once the algorithm could be shown to respond with accuracy, consistency etc. to novel data sets, one could accept its output. In this case, the stakeholder would be expected to trust the algorithm because it behaved in a way that was reliable.

An alternative (but logical) interpretation of this approach would be to assume that the human was able to run analysis on the same data used by the algorithm such that the output of these two analyses could be compared. While subject matter experts (SME) would be capable of analysing evidence or creating hypotheses for the situation which relate to the evidence, comparison of analyses could be problematic.

First, one purpose of applying AI / ML is to handle volumes of data that would be difficult for the human to analyse. Consequently, it is difficult to conceive of situations in which there could be a comparison of data at scale. This would mean that any comparison would be on partial data sets (either ones that were small enough for the human to be able to compute mentally or ones that could be analysed using other mathematical tools) such that the outputs could be compared. However, this raises problems of how one might sample data sets to produce smaller sets that were sufficiently representative and robust to permit comparison that could be generalised to the full data set. From the perspective of human intuitions, the SME might be able to interpret a subset of features and map this to prior experience (but not run the complete analysis). If the ML output accords with this intuition it could be accepted (and, conversely, if it contradicts the intuition, it could be rejected). The point is that the analysis performed (by ML and human expert) are not comparable in terms of selected features or scale of analysis.

Second, the analysis process applied by AI / ML is unlikely to mirror that of the SME, which means that comparison of process would not make sense. Consequently, any comparison would be at the level of outcome rather than process. As humans are adept are providing post-hoc rationalisation of output, there is a further problem that comparison could simply endorse the output of the AI / ML through the mere fact that the human was able to produce a plausible story for the data (even if this story bore no relation to the workings of the algorithm). This could be problematic because it precludes generalisation of the algorithm (or, worse, produces a human-understandable 'story' that is only loosely related to the output of the algorithm). As we shall see in *Section 5*, this problem of aligning

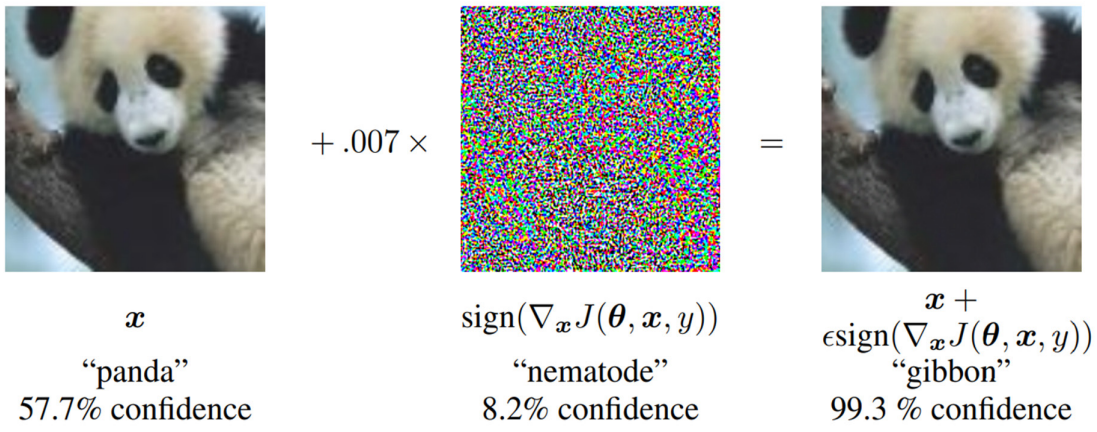


Figure 5. Confusing a panda for a gibbon, in which a small change causes disruption to weighting in a convolutional neural net [Szegedy et al., 2014]

the story to the underlying model can be applied to some of the proposed post-hoc approaches to support interpretability, such as LIME or SHAP.

Third, even if it were possible to directly compare analysis processes (i.e. algorithm with human expert), it is likely that the human would use additional experiences and information that were not available to the AI / ML. Consequently, the knowledge used in the analysis is unlikely to be identical. Users' experiences of ML and data science can have an impact on whether users trust the output of ML (Bhatt et al., 2020; Suresh et al., 2020; Kaur et al., 2020). Similarly, knowledge of the domain can influence trust in ML, e.g. people with lower knowledge of a domain might be more trusting of the ML (Nourani et al., 2020), presumably because they might not be able to critique the output (Merritt et al., 2015).

We conclude that explanation is not simply a function of the quality of the algorithm or the human stakeholder's ability to make sense of the algorithm. Rather, there are a host of additional factors that relate to human interaction with algorithms (*Table 2*). As Gunning (2019) proposed, in his discussion of DARPA's Explainable Artificial Intelligence (XAI) programme, "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners." From *Table 2*, the factors that might influence acceptance of AI / ML can also influence the way explanation is sought; in particular, *Table 2* raises concerns relating to understanding causality and the interpretation of risk. Both aspects imply a causal model that relates the knowledge used by the algorithm to the situation in which the knowledge applies, and which can project alternative states for the situation (either retrospectively to enable consideration of how situational states were caused or prospectively to enable consideration of the consequences of applying the decision). Of particular importance (to our discussion) is the notion of 'partner' in this quotation (in addition to 'trust' and 'understand') because a partner emphasises the social, interactive nature of the relationship between human and AI system. By implication, this social, interactive relationship can also refer to the ways in which explanations are provided. For the stakeholders involved in AI / ML, these causal models could involve information that goes beyond that which the algorithm is using (either because the human is considering situations that have not been modelled (in the form of what-if scenarios); or is considering information that is not part of the training data; or is considering second- or third-order consequences of the decision that lie outside the model parameters).

In this respect, even the apparently modest proposal to make the algorithm interactive could be highly problematic because there would be limits to the extensibility or flexibility of the algorithm (one could not, for example, simply add a bundle of new data to the algorithm and expect it to produce a solution, particularly if the new data was in formats that had not been defined in the original data set).

Trust	Human accepts that the technology can perform the activity without risk
Trustworthiness	Human has confidence that the algorithm will act as intended, even with new problem / data
Causality	Ability of algorithm to go beyond discovered pattern in the data to allow prediction of future 'effect'
Transferability	Ability of algorithm to operate beyond the constraints of its training data
Informativeness	Ability of algorithm's output to be assimilated into human decision-making
Confidence	Generalisation of stability and robustness
Fairness	Socially acceptable output
Accessibility	Interpretable by non-expert
Interactivity	Humans able to challenge or tweak model
Security	Protection of data and privacy

Table 2. Factors that might influence acceptance of AI / ML [from Arriete et al., 2020]

4. AUGMENTED INTELLIGENCE AND HUMAN DECISION-MAKING

Before discussing the nature of explanation, it would be useful to explore the role of augmented intelligence in human decision-making. Augmenting human intelligence is not simply the province of AI / ML but has a long history through which humans have sought to either offload cognitive activity or complement such activity by using physical media and machines. One perspective on this comes from theories of distributed cognition.

For example, Hutchins describes the ways in which the speed of an aircraft (1995) or a ship (2014) are calculated using a combination of devices, procedures, and people. On the ship, the devices might include equipment used to perform sighting of landmarks or fixed points at known distances against which the ship's movement can be timed, or nautical slide rules which support calculation of speed, or, in the aircraft cockpit, speed bugs in an aircraft to mark specific points on an airspeed indicator. The procedures might include the aircraft's flight plan that has been tailored to a specific runway in a specific airport, or the rules used to collect and analyse data for calculating speed.

The people might include a person on the ship's bridge performing the sighting, another person performing the timing, and a third person marking a map and performing the calculations. Across these different examples, Hutchins noted that there was not a single point at which speed was calculated as such, but rather the procedures allowed collection, collation, and analysis of data from different devices and people who were interacting in pursuit of a common purpose. From this perspective, AI / ML could simply be another 'device' or, depending on its complexity, it could be another person (or, at least, an intelligent agent who is part of the analysis team).

One of the reasons why explanation has been considered in terms of algorithmic transparency (see previous section) might be due to the device perspective. In this case, explanation becomes a matter of lifting the hood (on the algorithm) to allow humans to peer inside and appreciate the inner workings.

However, this assumes that the human has the knowledge and willingness to interrogate these workings. It also, perhaps more fundamentally, misses the point that AI / ML also introduces (often opaquely) modifications to the analysis procedures (in terms of the requirements for data formatting, the algorithms applied to these data, or the weighting applied to different features in the data etc.)

From this perspective, AI / ML is not simply an additional device which is introduced with no impact on procedures but, rather, causes a reallocation of functions such that some of the activities that had previously been performed by people or offloaded onto relatively passive devices that could store and process data, are now being performed by an (possibly) intelligent (but non-human) team member.

The relationship between humans and AI / ML can be considered in terms of levels of automation (Sheridan, 1992). At the one extreme, there is a level of no automation, in which humans perform all the cognitive and physical activity with no support, and, at the other extreme, there is a level of full automation, in which the computer performs all the activity with no human intervention (and, typically, with no information passed to the human regarding activity or outcome).

Between these extremes lie various levels which describe the allocation of activity or decision responsibility between human or computer. For example, lower levels might involve the human issuing

an instruction to the computer to perform an action and then vetting the output of the action, while higher levels might involve the computer autonomously beginning an action and then seeking approval from the human prior to completing the action. From this perspective, explanation could relate to several aspects of the action to be performed.

For example, if there is a choice of options then explanation would provide justification for the chosen action, or if there is the possibility of different outcomes then explanation would provide justification for the chosen action in light of expected or experienced outcome, or if there was a range of interpretations of the situation in which an action could be performed, then explanation could relate to the interpretation of the situation (e.g. in terms of the selection of features that were defined as salient). From this, explanation would have a variety of meanings because it is motivated by a variety of choices, actions, features, and outcomes.

For the human, making sense of the explanation could require knowledge of the situation in which the explanation was offered or the relationship between the explanation and background knowledge. In both cases, the receiver of the explanation (who we will refer to as the explainee) might not be able to interpret the explanation in terms of additional information that is not included in the content of the explanation. In human conversation, there is a wealth of strategies that the explainer uses to check the explainee's understanding of an explanation, but these strategies are less apparent in ML. Furthermore, the appropriateness of the explanation that is offered would depend on the situation in which the actions are performed as well as the knowledge, skills, and abilities of the team members. Interestingly, there is far more attention in the literature on providing explanation from automation to humans than vice versa.

In a series of studies, we have been exploring the likelihood that humans would accept recommendations or follow guidance offered by a computer. Comparing human decision-making with automated support, Morar and Baber (2017) shows that, after a few trials, people can adapt their response to the computer's accuracy (even when this is not explicitly stated). This was supported by a subsequent study (Baber et al., 2019).

This meant that, when the automated system had a reliability of 25%, people would rely on their own interpretation of the situation for the decision, but when the automated reliability was 81%, people would be more likely to rely on the computer. Overall decision accuracy was similar in both conditions (around 96%).

Interestingly, decision accuracy was higher when the computer responded first (in the experiment, the computer and the person took turns to propose the answer and then the human would submit a decision).

This was true even in the low-reliability conditions, which suggests that the computer's answer provided constraints on the options for the human to consider. We assume that the automation would provide the best answer it could, and the role of the human would be to interpret this in terms of a model of the task (in this case road-traffic monitoring). As such, the ability of the human to work with such a model, in terms of situation awareness, is crucial (Baber et al., 2019).

As users become more experienced in a task, so their ability to focus on key information improves. This was demonstrated by eye-tracking studies which show that experience leads to a solution involving three to four information sources (from a possible 12), and that the layout of the user interface can have a bearing on this activity (Starke and Baber, 2018). This search strategy can be modelled as optimal, i.e. a computer model (using a partially observable Markov decision process) that learned a policy to minimise search time while discovering the salience of information sources, produced similar behaviour to the humans in this task (Chen et al., 2017).

However, when the reliability of the automation decreases, search strategy changes. That is, only when the reliability approaches 100% will the human sample

a selection of sources (even when these are visually cued by the computer), and when the reliability falls to 90% or lower, people are more likely to check all the information sources (Starke and Baber, 2020). A computer model of this search strategy demonstrates that restricted sampling is optimal when reliability is known to exceed 96% (Acharya et al., 2019). Taken together, these studies suggest that people can be sensitive to the reliability of decision support; that a decision that is less than perfect will alter the information sampling strategy of users (and we believe this influences their decision-making); and that interpreting the output of the decision support requires understanding of the domain (and we believe that this domain knowledge is probably more important than knowledge of the algorithm).

We note (from discussion at workshops conducted as part of this project) that the allocation of function between human and ML is not simply a matter of the human interpreting and reacting to the output. Rather, there are many points at which humans are involved in the process flow on ML.

The following list is a generalised set of functions that humans might perform (and we would assume that these functions might be performed by different stakeholders with different appreciations of the underlying mathematics of the algorithm and / or different knowledge of the implications of the decision / outcome for operations):

- Identify and prepare relevant data sets
- Select algorithm
- Extract features and build model
- Refine analytical model (e.g. hyperparameters)
- Train model on test data
- Run and test on unseen data
- Refine model
- Run on new data

While the computer will run the analysis, there is much in this list which requires human intervention and these could introduce opportunities for human biases (intentional or otherwise) to affect performance, e.g.:

- through the selection of data (which might reflect social inequalities, sampling biases, lack of balance in the dataset, etc.)
- through the selection of hyperparameters and the tuning and refining of algorithms (which could skew the model to produce output is a good fit, e.g. with high precision and recall scores, but which could lead to socially unacceptable outcomes if the results are fed into policy).
- through the interpretation of good results (e.g. treating the results in terms of statistical models rather than consequences for policy and action).

For each of these functions (and for each type of stakeholder) the nature of explanation may differ. At its root, this means that explanation will not involve a one-size-fits-all solution but needs to respond to the nuances of differences in stakeholder or function.

When humans are provided with automated support for their decision-making, it is common to consider this relationship in terms of compliance, i.e. the extent to which the human will follow the recommendation of the computer, and reliance, i.e. the extent to which the human depends on the computer (Meyer, 2004).

High levels of compliance and reliance could mean that the human is unquestionably following the computer and is not able (or willing) to perform the task. In this case, the human could be said to be over-trusting of the computer.

At the other extreme, the human could be highly skeptical of the computer and unwilling to comply with its recommendations. From this perspective, the relationship between human and computer can be considered in terms of trust (itself a multi-factorial concept) and explanation would depend on the degree of trust between human and computer. Indeed, trust can

have a bearing on whether the explanation is believed (or whether there has been an error or obfuscation), or on whether the explanation is accepted (i.e. that explanation might be believed but the implications or consequences arising from the explanation might be doubted).

While much of the research on these issues have focused on trust expressed by human towards the computer, one could assume that many of the developments in AI / ML have been founded on mistrust of the human. That is, the algorithms have been developed with a view to minimising the risks of human failure (say, because data are too complex, too numerous, or too ambiguous for humans to be able to process accurately). Ironically, perhaps, the role of the humans in AI / ML systems is often to provide a defence against the failure of the algorithms.

From this perspective, explanation is related to the likelihood of disagreement (or, at least, the mismatch between the computer and human response to the data) to support agreement (or at least appreciation) of *how* the human or computer reached a particular conclusion. In one respect, this could relate to algorithmic transparency (with a focus on how the conclusion was reached). In another respect, this could relate to the setting of constraints on that conclusion, e.g. using counter-factual reasoning or variation of the weighting of features used in the analysis.

5. EXPLANATION FRAMEWORK DEVELOPED IN THIS PROJECT

From the discussions in the workshops, the consideration of explanation in machine learning (*Section 4)*, and the review of literature from human sciences (*Appendix A*), we have elaborated on the initial framework (*Figure 6*) to produce a definition of explanation.

We propose that an explanation, *E*, involves the set of features in a situation to which a person pays attention to, a means of defining the relevance, *R*, of these features, and a (potential) aim of influencing action, *A*. From this, an explanation is generated when two parties, X_1 and X_2, in a situation, *S*, agree the features to which each party pays attention to and agrees as relevant, with a view to altering X_2's version of *S* or *R* such that this could, potentially, lead to an action (*Figure 1*). To illustrate the framework, we use the following motivating example (we apply the framework to examples collected from workshops in the next section):

> *A hacker has obtained access to email accounts in your organisation and is sending scurrilous messages that appear to originate from people you work with. An investigation by your IT team, supported by an Intelligent Network Analysis System, results in a change to the management of the email system, and the problem is resolved. As a result of this, email users must create new passwords.*

We define 'situation' as a set of features that can be described symbolically as words, numbers, pictures, etc. That is, S = $\{f_i \ldots f_n\}$. A feature is some aspect of the situation to which people can attend. Individuals in a situation will ground their situation awareness, s_i, by attending to relevant features. That is, the attended set of features is a subset of all the features in the situation, $s_i \subseteq S$. These features imply (a) a string of causal reasoning that the other people are assumed to be able to perform and (b) to be sufficient to explain the situation. Features are assumed to be external, in

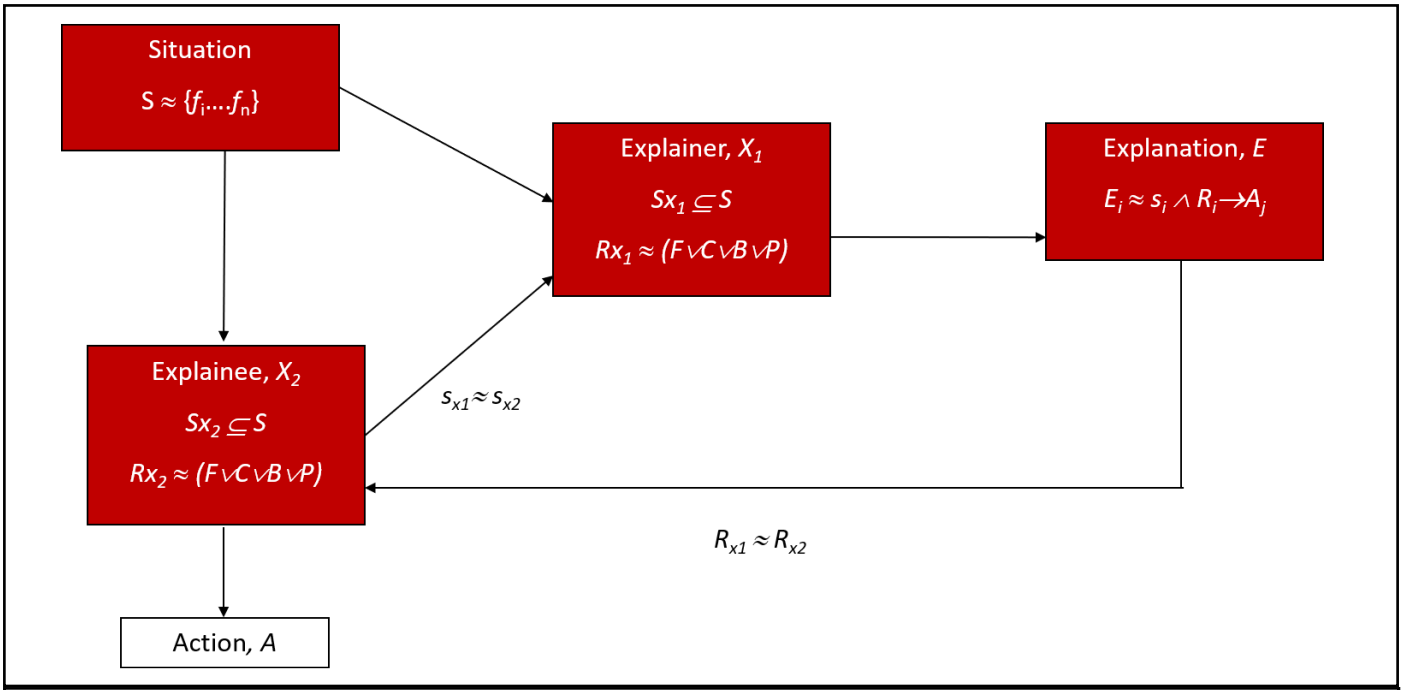


Figure 6. A framework for explanation

that anyone in a situation ought to be able to attend to the same features.

However, we accept that there will be situations in which some features might need to be inferred and might not be immediately accessible to all parties. In the example used in this paper (of resolving a hacked email system), the situation can be defined by features that include measures of network activity (defined in terms of normal and unusual), opportunities for resolving unusual activity, information to include in reports about network activity, and its resolution, etc.

The first challenge in explanation is to ensure that the set of features to which the explainer, X_1, pays attention to will overlap with the set of the explainee, X_2. That is, an initial goal in providing an explanation is to ensure that $s_{x1} \approx s_{x2}$. Notice that we do not need to assume that these sets are identical, only for there to be sufficient overlap (which is what concepts of common ground (Clark, 2015) would indicate, as will be expanded upon later in the paper). In part, this requires X_1 and X_2 to have overlapping feature sets (which might be particularly challenging if one or both parties are relying on internal, inferred features rather than external features).

The second challenge in explanation is to agree on what defines relevance. We propose (as a starting point) that relevance could be defined as:

- 'Clusters' identify which features co-occur but do not make predictions or draw inferences about feature relations.
- 'Beliefs' are based on prior experience that features co-occur and can predict consequences of specific features alterations.
- 'Policies' identify which features co-occur to allow actions.

There is, in this distinction between clusters, beliefs, and policy, an implied degree of strength of relevance. For example, messages across a network constitute a feature; a count of messages over time constitutes a cluster; whether the network is busy or not constitutes a belief; and responses to manage the network is a policy.

From this, we propose that an explanation, *E,* involves the set of features to which a person attends, a means of defining the relevance, *R*, of these features, and a (potential) aim of influencing action, *A*:

$$E_i = s_i \wedge R_i \rightarrow A_j \; [1]$$

where R = (C∨B∨P), A = action

From [1], an explanation is generated when two parties, X_1 and X_2, in a situation, *S*, seek to align the features to which each party attends, with X_1 seeking to alter the notion of relevance applied by X_2 knowing that this could lead to an action (*Figure 1*).

In this respect, selection of features from the situation involves a process analogous with Klein et al'.s (2007) Data-Frame concept of sense-making (where data is a set of features and frame is a form of relevance). For the hacking example, relevance could be framed in terms of a combination of features; the features could occur together (cluster); the features could imply a particular form of attack that the analysts have seen previously (belief); and analysts could have an agreed strategy for resolving the type of attack (policy). Notice also that a possible consequence of the explanation could be for x_2 to perform an action, *A*. The actions could be, for example, that X_2 acknowledges or accepts the explanation, that X_2 challenges the explanation or seeks further information, or that X_2 performs some task as a result of the explanation.

An explanation, in this case, ought to indicate how the features align to relevance. As in Lombrozo's (2010) hypothesis, different modes of cognition employ different modes of abductive reasoning, so that there is more than one type of explanation process. *Figure 9* suggests that initial alignment involves checking the features attended by x_1 and x_2. If these are not aligned, then the first-pass explanation might involve highlighting specific features, so that $s_{x1} \approx s_{x2}$. Where there continues to be uncertainty or misalignment, then

further action might be required to produce alignment across one or more type of relevance. Misalignment of belief could involve challenging the selection of features; misalignment of cluster could involve analysis using a different set of features; misalignment of policy could involve proposing a different strategy.

One of the reviews of this report has interpreted the framework (*Figure 6*) in terms of the realistic accuracy model of personality judgment (Funder, 1995). Drawing the analogy between making a judgment about personality traits, we can assume that an explanation assumes that *"...although truth indeed exists, there is no sure pathway to it. There is only a wide variety of alternative pathways, each of which is extremely unsure."* (Funder, 1995, p. 656). For Funder (1995), the accuracy of a judgment can be formalised as follows:

> Accuracy =
>
> [(relevance: of behavioural cues to a personality trait) x (availability: of cues to observers)]
>
> x
>
> [(Detection: of these cues) x (Utilisation: of these cues by the judge)]

In this case, the square brackets separate an environmental side of the equation (analogous to our notion of situation) from the perceiver side (analogous to our notion of explainer and explainee). Assuming that the explainer and explainee are individual judges, agreement on the cues (analogous to our notion of features) would be S1 ≈ S2 in our framework, and agreement on the utilisation of these cues would be R1 ≈ R2. In the following section, we present further examples of the framework to illustrate its application to the motivating example.

5.1 APPLYING OUR EXPLANATION FRAMEWORK TO HUMAN-HUMAN INTERACTION

Having defined elements of explanation, we can use these to produce some simple use-cases to illustrate the processes that might occur.

5.1.1 EXAMPLE 1: $SX_1 \approx SX_2$ AND $RX_1 \approx RX_2$

Let us assume that our two individuals, explainer, X1, and explainee, X2, have similar knowledge, training, experience, etc. In this instance, when both parties assume that Sx1 ≈ Sx2 and Rx1 ≈ Rx2, the need for explanation is negligible. However, when Sx1 ≠ Sx2, the individuals will need to resolve common ground, e.g. through agreement on which factors are relevant. If X1 ≈ X2, alignment could simply involve indicating a change in a relevant feature. We assume that there is honest signalling (Maynard Smith and Harper, 2003) in that the change in feature has occurred and that this change is relevant to the situation.

For example, the email traffic in the network might be unusually low for a Tuesday compared with previous weeks. In this case, X1 might draw the attention of X2 to this. However, if X2 does not recognise the relevance of this change in situation feature, then an explanation would involve X1 highlighting the change and presenting the associated belief as to its relevance. Here, the assumption is that (because X1 ≈ X2 is the equilibrium state) it should be possible for X2 to interpret the belief with minimal effort, i.e. X1 can highlight the relevant features and expect X2 to access a belief to determine relevance.

5.1.2 EXAMPLE 2: $SX_1 \approx SX_2$ AND $RX_1 \neq RX_2$

For people without similar backgrounds (i.e. X1 ≠ X2), alignment could be more effortful. Given the potential mutability of beliefs and clusters, it makes sense for explanation to focus initially on ensuring alignment on the set of features. As an initial move,

the focus on features would allow people to check their assumption that alignment exists or is possible or allow the explainer to provide the basis for the explainee to infer an appropriate cluster, belief, or policy.

5.1.3 EXAMPLE 3: $SX_1 \approx SX_2$ AND $RX_1 \neq RX_2$ AND $RX_2 \approx RX_1RX_1$

Assume that an experienced practitioner is providing a training example to a new apprentice. In this instance, the aim is not necessarily to create full equilibrium (that is, one does not expect the apprentice to know everything that the experienced practitioner knows). Rather, there is an expectation of a change in the knowledge of the apprentice towards a subset of the knowledge of the experienced practitioner: ΔR_{x2} $r_{x1} \subseteq R_{x1}$.

For this to occur, there is a need to establish that $S_{x1} = S_{x2}$. In this case, an explanation is (a) ensuring that X_2 attends to specific features to (b) encourage the knowledge of the relevance of these features to a policy, i.e. the operations that can be performed over the features. This could allow the apprentice to distinguish between two specific types of network attack.

5.1.4 EXAMPLE 4: $S_1 \neq S_2$ AND $R_1 \neq R_2$ AND $\Delta R_2 \approx R_1$ R_1 AND $A_2 = \Delta S_2$

While Example 3 emphasises explanation as an epistemic objective (to increase knowledge of X2), this might not be so important in an analyst-user interaction. In this case, the emphasis might be on ensuring that the user understands the situation (and the consequences of their actions on this): Ax2 = Δsx2. In other words, the emphasis is on motivating the user to change their password, etc.

It is arguable that this motivational objective is fully dependent on a change in knowledge (does it matter if the user does not understand the entire basis of the advice if they act as required?) In this case, the explanation would place more emphasis on the action, A, to perform and the constraints (and consequences) of this action. Change in knowledge would be required only as far as it supported this change in action, i.e. for X2 to have a productive understanding. Indeed, an aim would be for X2 to become their own explainer or to have the analyst-as-explainer replaced by another source, such as a leaflet, website, etc.

5.1.5 EXAMPLE 5: $SX_1 \neq SX_2$ AND $RX_1 \neq RX_2$

Assume that the resolution of the incident is communicated to the public by a newspaper story. In this case, the reader of the newspaper will have a third-hand account (via IT department to PR department to journalist) and only a partial view of the situation.

Furthermore, one can assume that the newspaper reader is unlikely to be an IT specialist, so would also have less technical knowledge. In this case, while the newspaper story might provide an explanation of the hacking (in terms of the broad nature of the event), it might lack sufficient detail to enable reconstruction of either situation or knowledge. If the newspaper reader wished to implement the fix to the problem (to prevent their own email account being hacked), then it is unlikely that the explanation here would be sufficient.

5.1.6 EXAMPLE 6: SX1 ≠ SX2 AND RX1 ≈ RX2

Assume that a formal report (product) is written following the incident. This product is consulted by other analysts (possibly in other organisations) who create their own report. In this case, X_1 is the report (rather than another person). One can assume some equivalence of knowledge (in terms of the training and experience of the analysts) but differences in their access to the situation. In a sense, this sequence of formal reports is analogous to research on transmission chains (Bartlett, 1932). As information passes through a transmission chain, so it loses redundancy and becomes more focused (Kempe et al., 2019).

This might be the result of the formal structures imposed by the style of reports; of the way people share information; or of a desire to focus on relevant

information. A consequence of this might be that fewer and fewer of the situation factors become shared – until, somewhere down the line, a reader might challenge the report because it does not correspond to their interpretation of the situation. At this point, there might need to be communication between this reader and the report's originator, with a view to establishing the relevant factors of the situation.

When the X_2 does not agree with the explanation provided and / or does not understand it, it is important to consider how the X_2 decides that an explanation is not sufficient for their goals and understanding. Miller (2019) referred to this as 'explanation evaluation' and concluded that the most important and agreed upon criteria are: probability, simplicity, generality, and coherence with prior beliefs.

So, in our hacking example, an X_2 is most likely to accept an explanation that a) is consistent with their beliefs about email hacking (coherence); b) includes less causes but explain more of the events (simplicity, generality); and c) that a particular type of attack is the true cause of the observed features, e.g. the influx of unsolicited mail (probability).

Note that the simple statistical relationship (cluster) between a particular type of attack and the quantity of unsolicited mail is not sufficient explanation – causes are desired to explain events (Halpern and Pearl, 2005a). As mentioned earlier, while a true / likely cause is an attribute of a good explanation, to say that the most probable cause is the best explanation would be incorrect (Hilton, 1996).

5.2 APPLYING OUR EXPLANATION FRAMEWORK TO HUMAN-AGENT INTERACTION

Having developed a framework for human-human explanation and provided some illustrative examples, we consider how these explanation types might apply to human-agent interactions.

Logistic Regression	Independent Variables	Predictors and Interactions	Regression model for outcome	
Decision Trees	Nodes	Causal relations	Cause-effect relations as regression models	
K-nearest Neighbour	Labelled data points in n-dimensions	Defined by number (k) of neighbours	Clusters will be mutually exclusive	
Support Vector Machines	Variables	Classes	Likelihood of event (from classification)	Optimal division between classes

Table 3. Aspects of relevance for ML algorithms

5.2.1 EXAMPLE 1: S1 ≈ S2 AND R1 R2

Recommender systems might inform their users of the specific features that inform the recommendation, e.g. a word cloud taken from a movie's reviews (Gedkili et al., 2014), or a histogram of ratings of a move by similar users (Herlocker et al., 2000).

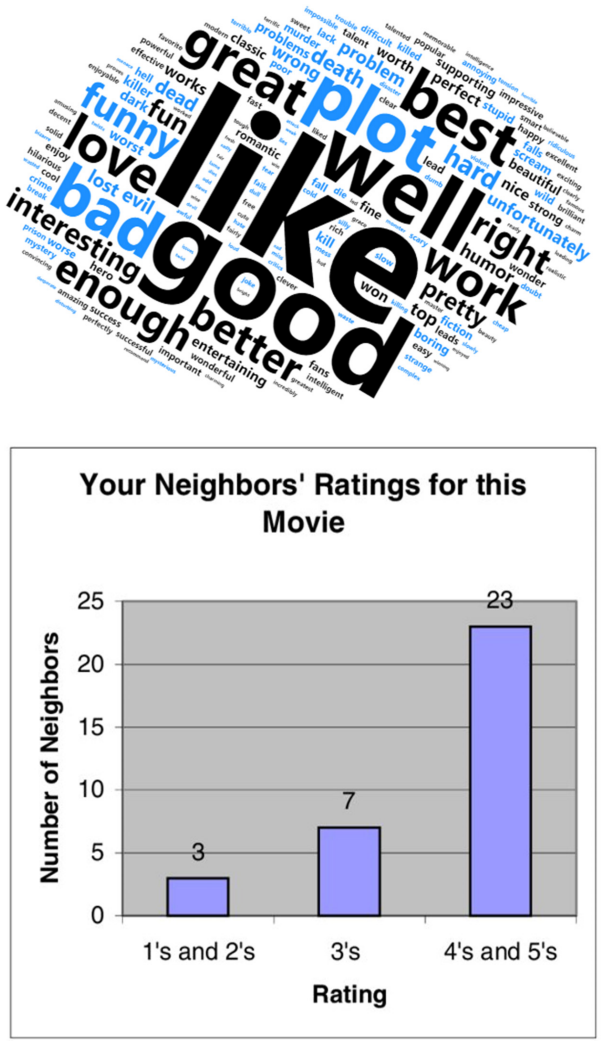


Figure 7. Word cloud (Gedkili et al., 2014) and MovieLens (Herlocker et al., 2000)

In our terminology, relevance is presented in the form of a cluster. Other recommender systems present a comparison of items. ExpertClerk (Shimazu, 2002) would offer a recommendation in terms of trade-offs of specific features, e.g. "This necktie is more expensive but is made of silk. That one is cheaper but is made of polyester." Here, objects are compared on two features

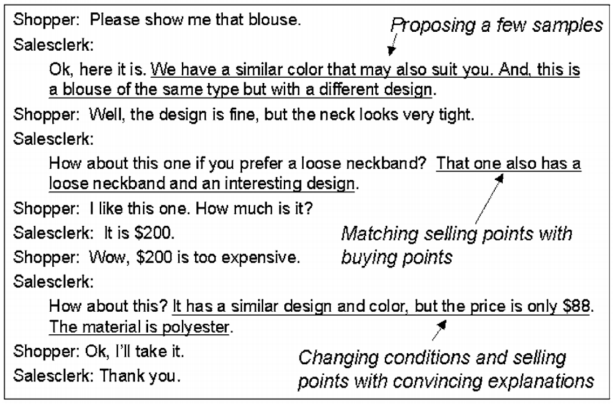


Figure 8. ExpertClerk dialogue with shopper (Shimazu, 2002)

and the trade-off is presented in terms of what we consider is a belief that could be discussed.

5.2.2 EXAMPLE 2: SX1 ≈ SX2 AND RX1 RX2

In most applications of ML, identifying a cluster does not involve indication of a belief. We note that the word belief is used in some forms of ML but has quite a different meaning to the way we used it. For instance, in a Bayesian Belief Network (BBN) situation, features are arranged in a network (*Figure 9*). Connections within this network are defined by probabilities and altering these probabilities produces different output. For BBN, belief is the probabilistic weighting of these connections. However, when a belief network becomes large, it can be difficult to read or to appreciate the relations that are being expressed. Techniques such as BayesPiles (*Figure 10*) can help in visual simplification of these networks.

From our perspective, the weighting of connections is, at best, a cluster and more likely simply a set of features (as far as the human decision-maker is concerned). This means that the BBN does not express a belief about its outcome, i.e. it does not offer a plausible, generalisable frame in which to make sense of the connections between features or account for what might happen if features are missing.

In other words, there is no underlying model (outside the data) that would allow prediction from the cluster.

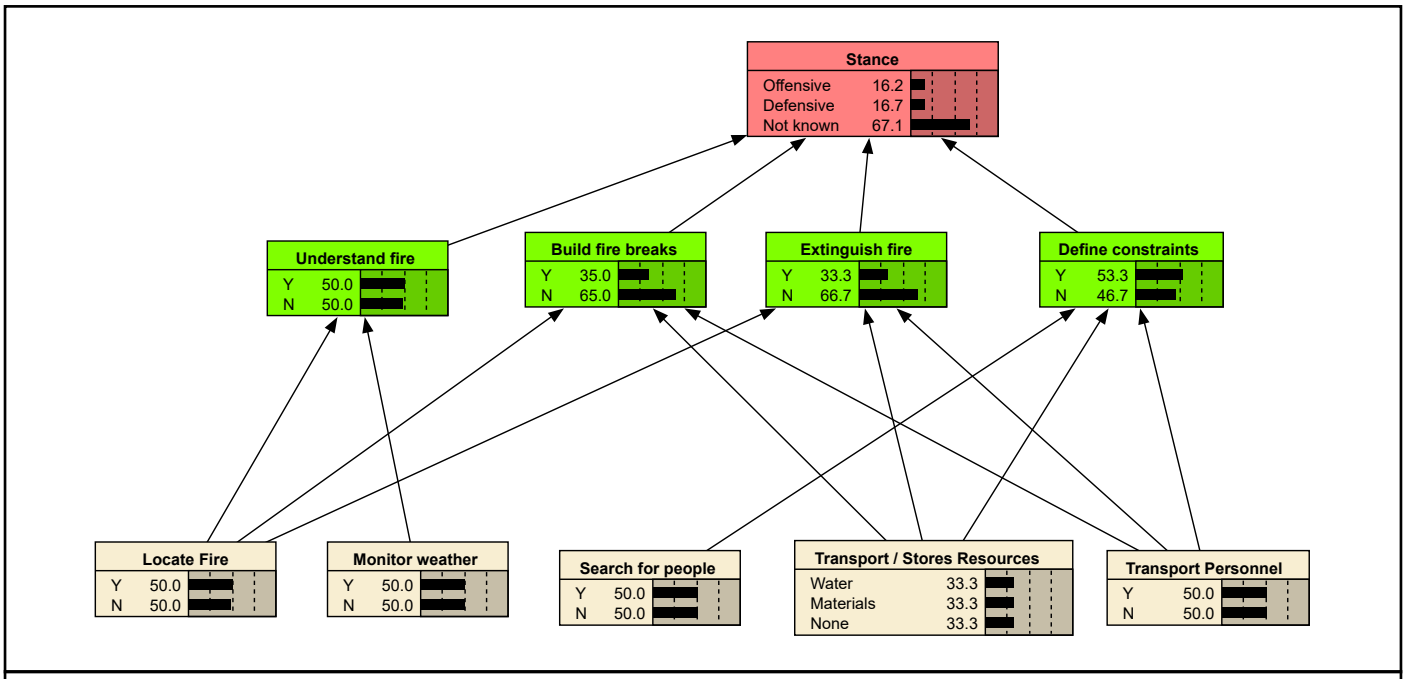


Figure 9. Sample output from NETICA

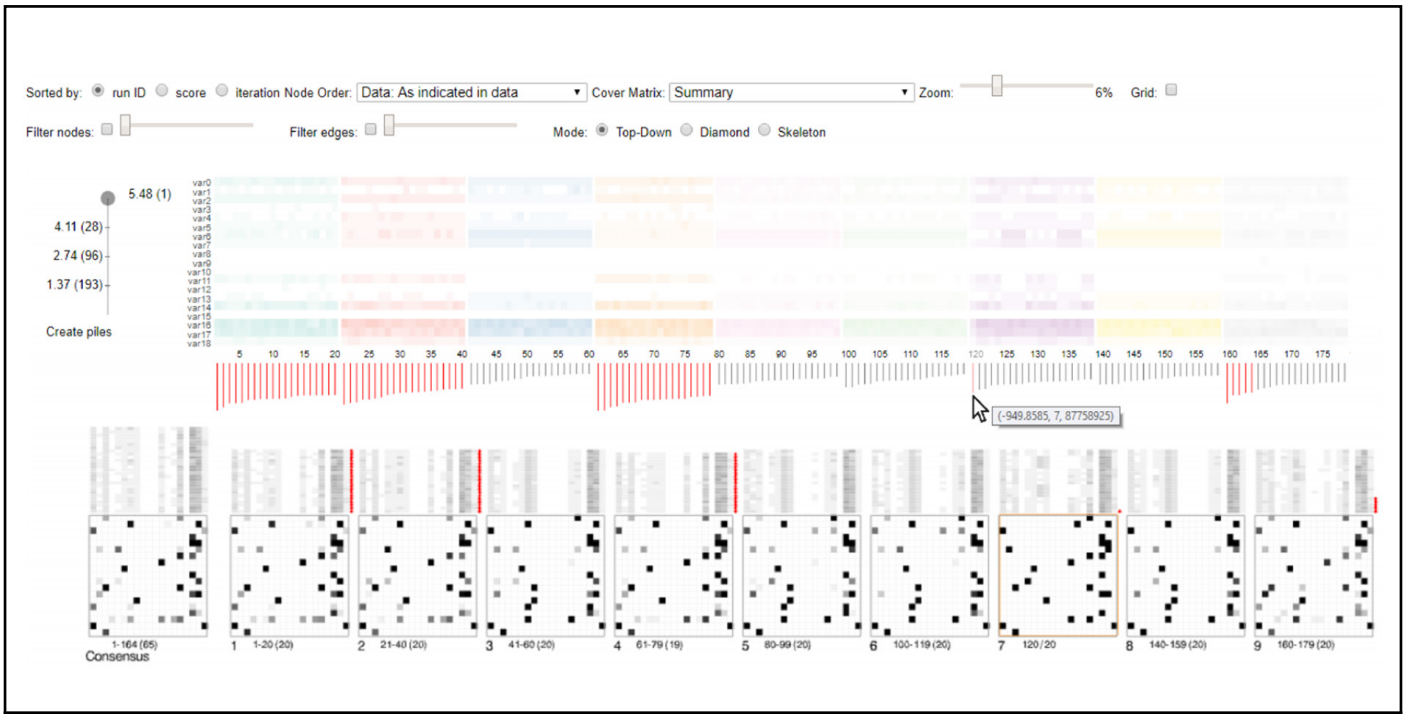


Figure 10. Example visualisation of BayesPiles [Vogogias et al., 2017]

In this case, while there might be an intention of aligning the features that the ML has used with those that the human can interpret, such that $S_1 \approx S_2$, it is not possible to align notions of relevance. However, users might assume belief from the output of ML, e.g. either anthropomorphising the process by which an outcome has been reached or assuming that counter-factual reasoning would be possible by modifying the features that the ML uses. For ML to provide something that might be interpreted as beliefs, it is possible to apply techniques for association rule mining, which seek to discover dependencies between features in ways that are more amenable to generalisation than clusters would allow (Altaf et al., 2017).

5.2.3 EXAMPLE 3: SX1 ≈ SX2 AND RX1 RX2 AND ΔRX2 ≈ RX1 RX1

Educational technologies provide personalised and adaptive environments to support learning (Dawson et al., 2010). In these systems, learners are provided with situations in which they review material (features) to answer questions (actions) and their performance on the questions will impact on their progression through the set of material, i.e. learners who make mistakes or show misconceptions will be provided with more material of similar content and more questions of similar difficulty.

Furthermore, some systems seek to align attention to information (relevant features) and knowledge (understanding of operations over the information) through open learner models that compare to the learner's understanding (based on their answer to specific questions) and their confidence in giving these answers.

5.2.4 EXAMPLE 4: S1 ≠ S2 AND R1 R2 AND ΔR2 ≈ R1 R1 AND A2 = S2

Technology-mediated 'nudging' (Caraban et al., 2019) create 'choice architectures' that present alternative actions to decision-makers in ways that are intended to support positive changes in behaviour. These technologies encourage or discourage behaviours that might have impact on the user's wellbeing. These technologies might remind the user of specific consequences of their actions, suggest alternative actions, or emphasise social desirability of the consequences. In our terms, the focus is on action, through highlighting relevant beliefs.

5.2.5 EXAMPLE 5: SX1 ≠ SX2 AND RX1 RX2

Explainable AI can be defined as a situation in which the explainer (the AI) attends to a different set of features to those used by the explainee, and the definition of relevance used by the two parties does not align. In deep (or reinforcement) learning, the AI seeks to discover a policy by which it can optimise reward (say, success in playing a game) by performing actions in specific situations. Post-hoc analysis of the AI performance (e.g. in the form of gradient-based saliency plots, *Figure 19*) could allow the person to infer the features that the AI might have been using, i.e. $S_{x1} \approx S_{x2}$. However, it is not so easy to discern how the features were defined as relevant or even whether the AI actually made use of these features. Combining a host of outputs, from the application of different algorithms, could allow the analyst to 'compare and contrast' the relevance of different features in terms of policy. (*Figure 11*).

5.2.6 EXAMPLE 6: SX1 ≠ SX2 AND RX1 RX2

Argumentation technology (Reed et al., 2017) combines a computer model of reasoning towards conclusions (arguments) with an interface that allows users to explore the structure of these arguments. We assume that the features, or relevance, offered by parties in an argument might not align. Through argumentation, parties identify points of similarity and difference, e.g. features to emphasise or notions of relevance. User interfaces for argumentation visualise the set of features drawn upon by an argument and their relations (which we would call beliefs). The user could then explore the effect of adding or removing features or changing relations, which could be particularly useful for counter-factual reasoning (Guidotti et al., 2019).

Appendix B develops these ideas with an application to one of the use-cases developed during the workshops.

5.3 HOW MIGHT THE FRAMEWORK ACCOUNT FOR BREAKDOWN AND REPAIR IN EXPLANATIONS?

From our explanation framework, we can also consider how breakdown might arise and what might be required to enable repair.

For human-human interaction, one can imagine ways in which the explainer and explainee will engage in conversation to check their selection of features or definition of relevance, and to ensure that the definition is shared and understood sufficiently to enable an action to be performed.

When the explainer is an algorithm, the process of aligning features or defining relevance can be more difficult. Advances in visual analytics (which allow humans to explore data sets and clusters) could offer opportunities to ensure alignment of features if the algorithm is capable of recalculation when the human adds or removes features.

For some algorithms, the latter might be difficult (because removing features unbalances the data or redistributes the correlations between features). It might be useful to produce multiple runs of the algorithm, using a leave-one-out approach to the sets of features, to enable the exploration of sensitivity of the algorithm to different combinations of features. But such exploratory data analysis (or 'fishing') is not always good practice and can reinforce biases or misconceptions held by analysts, particularly if their knowledge and experience of the domain is limited.

It might be preferable to aim for explanation type 3, in which the human analysts, through education and

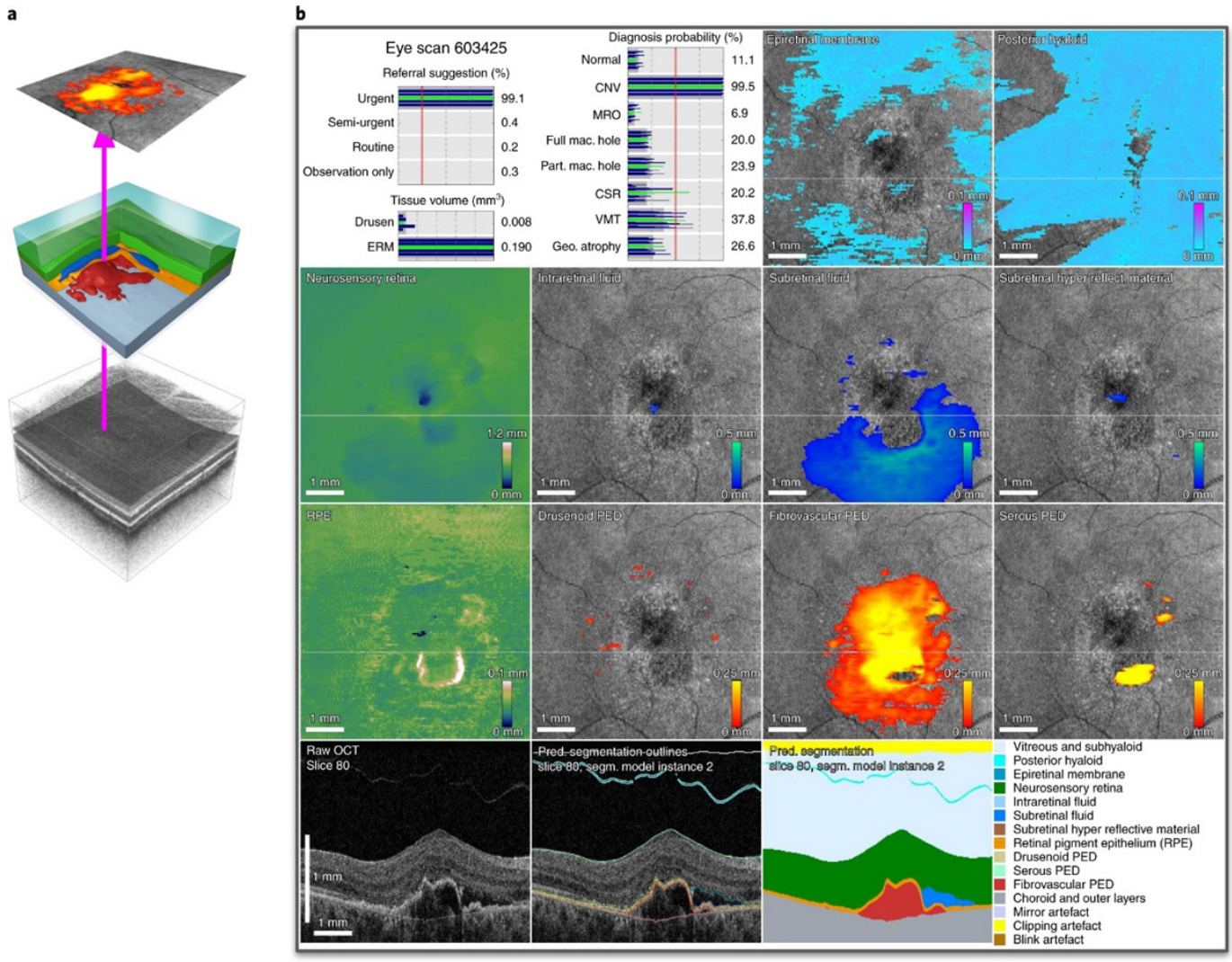


Figure 11. Multiple views of deep learning outputs for retinal diseases diagnosis [De Fauw et al., 2018]

feedback from the algorithm, are able to understand the definition of relevance that the machine is applying. This looks very much like the approaches to intelligibility and transparency of algorithms discussed in *Section 3*.

However, without understanding the features (i.e. without a sufficient situation awareness) simply understanding the algorithm is insufficient to appreciate the implications of its recommendation and output. In this respect, explanation type 4 (i.e. knowing how complying with the recommendation will affect the situation) becomes essential. This corresponds to level 3 situation awareness (*Section A.5*).

Since we have conceptualised explanation as an interactive process, it is important to consider how to challenge an explanation that explainees do not understand or do not agree with. To dispute an explanation given, an explainee should use the following methods:

- Deduction – reasoning from more premises to reach a logically certain conclusion.
- Induction – making broad generalisations from specific observations / knowledge.
- Abduction – inference from evidence to the best explanation for a given conclusion
- Argument by analogy – perceived similarities to infer further similarity that has yet to be observed
- Reductio ad absurdum – either disprove a statement by showing the result would be absurd, or to prove one by showing that if it were not true, the result would be impossible.

	Explanation Type	Breakdown	Repair
1	S1 ≈ S2 and R1 ≈ R2	This assume that both parties agree features to extract from the situation and the appropriate definition of their relevance. Breakdown in either of these results in types 2, 5 or 6	Either seek agreement of feature set or alignment on definition of relevance
2	Sx1 ≈ Sx2 and Rx1 Rx2	The feature sets are aligned but there is not agreement on their relevance	Seek alignment on definition of relevance
3	Sx1 ≈ Sx2 and Rx1 ≠ Rx2 and ΔRx2 ≈ rx1⊆Rx1	The feature sets are aligned and the aim is to shift the definition of relevance held by x2, e.g. through teaching	Guide x2 to acquire understanding of relevance. Check that x2 can apply the new R.
4	S1 ≠ S2 and R1 ≠ R2 and ΔR2 ≈ r1⊆R1 and A2 = s2	As 3, but with the added aim of encouraging action by x2.	As 3, plus encourage action. Check that x2 understands the goal of the action. Check that x2 performs the action acceptably.
5	Sx1 ≠ Sx2 and Rx1 ≠ Rx2	See 1	See 1
6	Sx1 ≠ Sx2 and Rx1 ≈ Rx2	See 1	See 1

Table 4. Mapping breakdown and repair to explanation types

These discussion methods can be used in many ways to suit the situational factors that constrain the explanation. For instance, the explainee could use a simple question to challenge the logic used in the original explanation.

It is important to not stray far from these types of arguments because other types of debating techniques can stem from illogical reasoning or from judging the explainer's reasoning based on facts that are unrelated to the explanation given.

For example, an ad hominem argument is one that is directed against a person rather than the position they are maintaining. Challenging explanations in this way can lead to unproductive discussion, a loss of trust between parties, and decisions that result in unsatisfactory action further down the line.

Example 1:

Person 1:
I think we should use approach X to tackle the email hacking issue because it has been an effective method in other similar hacking cases, like with company Y.

Person 2:
(thinks person 1 is too young and inexperienced to know this for sure so seeks out another opinion from someone they think will disagree with person 1. This results in confirmation bias).

Furthermore, by challenging an explanation productively, it gives the explainer the opportunity to see the fault in their original reasoning so that the same mistake will not be made in the future.

Example 2:

Person 1:
I think we should use approach X to tackle the email hacking issue because it has been an effective method in other similar hacking cases, like with company Y.

Person 2:
Yes, but company Y used a different firewall to the company we're looking at now. Do you think approach X can still solve this?

Person 1:
I hadn't thought of that, I'll remember to check that in the future.

6. GUIDELINES FOR EXPLANATIONS

In an overview of the impact of ML and related algorithms on human decision-making, Spiegelhalter (2020) offers some rules of thumb that could be asked of any algorithm (or, indeed, any decision-support automation):

- Is it any good when tried in new parts of the real world?
- Would something simpler, and more transparent and robust, be just as good?
- Could I explain how it works (in general) to anyone who is interested?
- Could I explain to an individual how it reached its conclusion in their particular case?
- Does it know when it is on shaky ground, and can it acknowledge uncertainty?
- Do people use it appropriately, with the right level of scepticism?
- Does it actually help in practice?

Aim of ML algorithm	What can you ask the data used by the algorithms?	What can you ask about the performance of the algorithms?
Classification	Sources of data Number of samples How were the data cleaned and otherwise prepared?	Correct versus false classifications {sensitivity; specificity; Precision; F1} Confusion matrix
Clustering	Sources of data Number of samples How were the data cleaned and otherwise prepared? Were the data labelled (by hand) or were clusters discovered automatically? Number of categories reported	Separation of clusters Tightness of clustering
Regression	Sources of data Number of features How were the data cleaned and otherwise prepared?	Correlation Fitting error
Dimensionality Reduction	Sources of data Number of samples How were the data cleaned and otherwise prepared? Number of dimensions reported	Under / Over-fitting Feature selection Filtering Feature extraction

Table 5. Types of ML algorithm and topics

These rules of thumb are not concerned with the workings of the algorithm so much as the context in which the algorithm is applied. They require that anyone deploying the algorithm should be able to appreciate how it contributes to the real world. In particular, the analyst should be able to understand the output of the algorithm sufficiently to explain how it arrived at this output and to interpret the output with sufficient scepticism to be able to respond to questions or challenges to its output. In answering any of these questions, there is a need to provide an explanation of the algorithm's performance, its output, and the impact of this output.

In its guide to 'Explaining decisions with AI: Part 1', the Turing Institute offers four guiding principles:

- Be transparent
- Be accountable
- Consider context
- Reflect on impacts (fairness, safety, and performance)

For this report, the questions point to two essential features of an explanation. The first is that explanations are situation specific, and the second is that explanations are tailored to the knowledge and experience of the explainee. From this, we propose that the guiding principles listed above can be applied to explainer and explainee as human actors in an organisation, and to the design, development, deployment, and use of AI / ML (in that AI / ML could be expected to be given the role of explainer).

A broad classification of AI / ML is given in *Table 5* (Appendix B provides an overview of how AI / ML methods relate to concepts of explanation). Without going into detail on how the 'aims of the machine learning algorithm' are defined, it is sufficient to recognise that these define broad types of algorithm. In some instances, the aim is to identify patterns (clusters) or associations (regression) in data sets. These clusters could then be used for classification (i.e. to identify a new piece of data in terms of known classes or clusters). Alternatively, the aim might be to reduce a very large data set to a small number of dimensions (usually, each dimension is defined by a cluster of features).

In the second and third columns, we suggest topics that a non-specialist might ask about the data that the algorithms use and about how these algorithms can be evaluated. These topics could be used for a first-pass discussion of what the algorithm will be used for, what data it might use, and how its performance could be judged.

Taking the notion of topics in *Table 5* a little further, we can propose a checklist that the non-specialist might use to ask questions about the potential use of AI / ML in their activity. This is shown in *Figure 12* and is intended as an aide memoire rather than a formal procedure.

From our review of the literature across different disciplines, we define common themes that contribute to the concept of explanation. In some places, this has meant merging terms for convenience. For instance, plausibility and likelihood were deemed synonymous in their respective contexts. Furthermore, some criteria were grouped under an overarching category if they related to a broader aspect of explanation. For example, plausibility and probability were aspects of explanation that referred to causes. In order of importance, the criteria are:

1. **Explanations should include relevant causes**
   Explanations should relate to beliefs in the relationship between features of a situation and the causes that can directly affect the event being explained (probability) or can explain most of the event (explanatory power); are plausible (construct validity); and if the cause was instigated by a person, deliberative.
2. **Explanations should include relevant features**
   Explanations should relate to the key features of the situation and the goals of the explainer and

explainee.

3. **Explanations should be framed to suit the audience**
   Explainers should fit the explanation to suit the explainee's understanding of the topic and what it is they wish to gain from the explanation (their mental model and goals).
4. **Explanations should be interactive**
   Explainers should involve explainees in the explanation.
5. **Explanations should be (where necessary) actionable**
   Explainees should be given information that can be used to perform and / or improve future actions and behaviours.

From the explanation framework that we offer, we propose guidelines relating to the dialogue between explainer and explainee:

- Seek alignment in features used in the explanation.
- The explainer should provide a clear and concise account of the features that are used in the explanation.
- The explainee might sketch the features expected before seeking the explanation. For example, Belgian police officers speak of 'think steps' for specific categories of crime – these represent the high-level activity that they would expect to perform in investigating a specific crime, with the associated classes of information that the activity

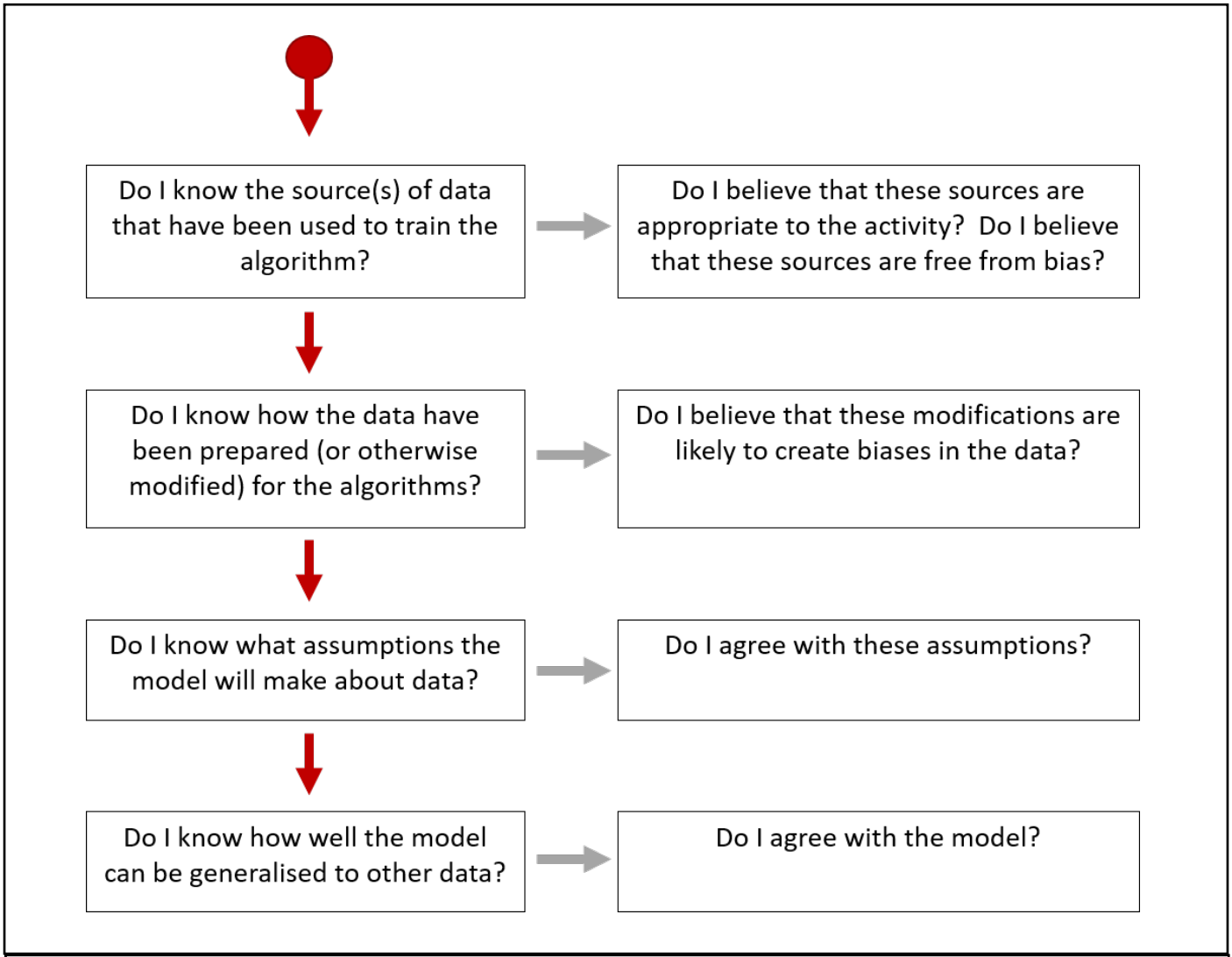


Figure 12. Flowchart for questions to ask of the AI / ML algorithm

involves (Hepenstal et al., 2019). In this way, the explainee could compare expected with observed features.

- The explainer and explainee should seek alignment on the features that are used. If there is discrepancy, this could either indicate differences in access to data or differences in interpretation of the features.
- Clarify the definition of relevance used in the explanation.
- We distinguished between different types of relevance – feature, cluster, belief, policy – that, while primarily derived from computing, can be related to the other domains that we have considered. Broadly, these information types imply a level of processing from the specific features that are recognised in a situation to the groupings of features into clusters, to the covering law or rules by which the clusters are formed and the implications of these rules for action (which we term belief), to the overarching regulation (we call policy) that governs acceptable actions in that context:

1. Defining features – as noted above, the choice of features used in an explanation will depend on their availability and their weighting to the analysis.
2. Defining clusters – for algorithms that seek to discover clusters in data, the visual display of the output might be sufficient for the explainee to make sense of the features and the clusters. However, it is important for the explainee to appreciate that the clusters are entirely empirical; changing the set of features (or the coefficients which define the clusters) can produce different results. The algorithm has no underlying belief or expectation as to why such features arise but, likely, the explainee might impose beliefs.
3. Defining beliefs – for algorithms that apply rules to the analysis of the data (e.g. in terms of maximising a reward function), the explainee should be able to not only understand how the beliefs apply to the data but also to challenge these beliefs, e.g. in the form of foils (counter-factual examples or additional features). Equally, algorithms could present foils to explainees as a means of challenging assumptions and bias in the humans.
4. Defining policy – in this report, policy relates to the actions which arise from an algorithm's recommendation. In this respect, understanding the second- and third-order consequences of accepting and acting upon an algorithm's output (or declining to follow this) plays an important role in explanation. It could also allow the algorithm to be modified or retrained to minimise the risk to unacceptable policy implications.

READ MORE

Acharya, A., Howes, A., Baber, C. and Marshall, T. (2018) Automation reliability and decision strategy: a sequential decision model for automation interaction, *Proceedings of the Human Factors and Ergonomics Society 2018 Annual Meeting,* Santa Monica, CA: HFES, 144–148

Amir, D., Amir, O.: Highlights: summarizing agent behavior to people. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1168–1176. International Foundation for Autonomous Agents and Multiagent Systems (2018)

Anjomshoae, S., Najjar, A., Calvaresi, D., Fr¨amling, K.: Explainable agents and robots: results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019)

Arias, A. M., Davis, E. A., Marino, J. C., Kademian, S. M., & Palincsar, A. S. (2016). Teachers' use of educative curriculum materials to engage students in science practices. *International Journal of Science Education*, *38*(9), 1504–1526

Aronson, J. L. (1971). On the Grammar of 'Cause'. *Synthese*, 414–430

Baber, C., Morar, M.S. and McCabe, F. (2019) Ecological interface design, the proximity compatibility principle, and automation reliability in road traffic management, *IEEE Transactions on Human-Machine Systems, 49,* 241–249

Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: human considerations in context-aware systems. Human–Computer Interaction, 16(2–4), 193–212

Bird, A. (1999) Explanation and Laws, *Synthese, 120,* 1–18

Borgo, R., Cashmore, M., Magazzeni, D.: Towards providing explanations for AI planner decisions. arXiv preprint arXiv:1810.06338 (2018)

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science education*, *95*(4), 639–669

Broekens, J., Harbers, M., Hindriks, K., van den Bosch, K., Jonker, C., Meyer, J.-J.: Do you get it? User-evaluated explainable BDI agents. In: Dix, J., Witteveen, C. (eds.) MATES 2010. LNCS (LNAI), vol. 6251, pp. 28–39. Springer, Heidelberg (2010)

Chakraborti, T., Sreedharan, S., Grover, S., Kambhampati, S.: Plan explanations as model reconciliation. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 258–266. IEEE (2019)

Chen, J.Y., Procci, K., Boyce, M., Wright, J., Garcia, A., Barnes, M.: Situation awareness-based agent transparency. Technical report, Army Research Lab Aberdeen Proving Ground MD Human Research and Engineering (2014)

Chen, X., Starke, S.D., Baber, C. and Howes, A. (2017) A cognitive model of how people make decisions through interaction with visual displays, In *CHI'17: Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, New York: ACM

Clark, H. H., & Brennan, S. E. (1991). *Grounding in communication.* In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (p. 127–149). American Psychological Association

Clark, H.H. (1991) *Using Language,* Cambridge: Cambridge University Press

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction, 18(5), 455

De Fauw, J. et al., 2018, Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 24, 1342–1350

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, *93*(3), 283

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, *23*(4), 517–542

Dowe, P. (1992). Wesley Salmon's process theory of causality and the conserved quantity theory. *Philosophy of science*, *59*(2), 195–216

Endsley, M.: Measurement of situation awareness in dynamic systems. Hum. Factors 37, 65–84 (1995)

Fair, D. (1979). Causation and the Flow of Energy. *Erkenntnis*, *14*(3), 219–250

Floyd, M.W., Aha, D.W.: Incorporating transparency during trust-guided behavior adaptation. In: Goel, A., D´ıaz-Agudo, M.B., Roth-Berghofer, T. (eds.) ICCBR 2016. LNCS (LNAI), vol. 9969, pp. 124–138. Springer, Cham (2016)

Fox, M., Long, D., Magazzeni, D.: Explainable planning. arXiv preprint arXiv:1709.10256 (2017)

Fromkin, V. A. (1973). *Speech Errors as Linguistic Evidence*. The Hague, Netherlands: Mouton

Funder, D.C. (1995) On the accuracy of personality judgment: a realistic approach, *Psychological Review, 102,* 652–670

Greydanus, S., Koul, A., Dodge, J. and Fern, A. (2018) Visualizing and understanding Atari agents, https://arxiv.org/abs/1711.00138v5

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill

Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, *40*(2), 44–58

Hacibeyoglu, M. and Ibrahim, M.H. (2018). The Effect of Over-sampling and Undersampling Techniques in Medical Datasets)

Harbers, M., Bradshaw, J.M., Johnson, M., Feltovich, P., van den Bosch, K., Meyer, J.-J.: Explanation in human-agent teamwork. In: Cranefield, S., van Riemsdijk, M.B., V´azquez-Salceda, J., Noriega, P. (eds.) COIN -2011. LNCS (LNAI), vol. 7254, pp. 21–37. Springer, Heidelberg (2012)

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science*, *56*(4), 843–887

Halpern, J. Y., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science*, *56*(4), 889–911

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related reply to Baars et al. (1975). *Journal of Memory and Language*, *52*(1), 58–70

Harley, T. (2008). *The Psychology of language*. 4th ed. New York: Psychology Press, pp.397–451

Hellström, T., Bensch, S.: Understandable robots-what, why, and how. Paladyn J. Behav. Robot. 9(1), 110–123 (2018)

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of science*, *15*(2), 135–175

Hepenstal, S., Wong, B.L.W., Zhang, L. and Kodogoda, N. (2019) How analysts think: a preliminary study of human needs and demands for AI-based conversational agents, *Proceedings of the Human Factors and Ergonomics Society 2019 Annual Meeting,* Santa Monica, CA: HFES, 178–182

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In Proceedings of the 2000 ACM conference on Computer supported cooperative work (pp. 241–250). ACM

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, *2*(4), 273–308

Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems*, *32*(3), 68–73

Hoffman, R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018). Explaining explanation, part 4: a deep dive on deep nets. *IEEE Intelligent Systems*, *33*(3), 87–95.

Holzinger, A., Carrington, A. and Müller, H. (2020) Measuring the Quality of Explanations: the Systems Causability Scale (SCS), Künstliche Intelligenz, 34, 193–198

Holzinger, A.,Plass, M., Kickmeier-Rust, M., Holzinger, K., Crisan, G.-C., Perita, C.-M., and Palade, V. (2018) Interactive machine learning: experimental evidence for the human in the algorithmic loop, *Applied Intelligence, 49*, 2401–2414

Hume, D. (2000). An enquiry concerning human understanding. In *Seven Masterpieces of Philosophy* (pp. 191–284). Routledge

Hutchins, E., 1995. How a cockpit remembers its speeds. *Cognitive science*, *19*(3), pp.265–288

Hutchins, E., 2014. The technology of team navigation. In *Intellectual teamwork* (pp. 205–234). Psychology Press

Jaspars, J. M., & Hilton, D. J. (1988). Mental models of causal reasoning.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Vaughan, J.W. (2020) Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, *CHI 2020,* New York: ACM, paper 92

Kelemen, D. (2019). The Magic of Mechanism: Explanation-Based Instruction on Counterintuitive Concepts in Early Childhood. *Perspectives on Psychological Science*, *14*(4), 510–522

Kim, B., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). arXiv preprint arXiv:1711.11279 (2017)

Klein, G.A. (1989) Recognition-primeddecisions," in Advances in Man-Machine Systems Research, ed. W.B. Rouse, Greenwich, CT: JAI Press, Inc., 47-

Klein, G. (2018). Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems*, *33*(2), 83–88

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems*, *21*(4), 70–73

Leddo, J., Abelson, R. P., & Gross, P. H. (1984). Conjunctive explanations: When two reasons are better than one. *Journal of Personality and Social Psychology*, *47*(5), 933

Leite, R.A., Gschwandtner, T., Miksch, S., Kriglstein, S., Pohl, M., Gstrein, E. and Kuntner, J., 2017. Eva: Visual analytics to identify fraudulent events. *IEEE transactions on visualization and computer graphics*, *24*, 330–339

Lewis, D. (1974). Causation. *The journal of philosophy*, *70*(17), 556–567

Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)

Livengood, J., & Sytsma, J. (2020). Actual causation and compositionality. *Philosophy of Science*, *87*(1), 43–69

Lomas, M., Chevalier, R., Cross, E.V., Garrett, R.C., Hoare, J., Kopack, M.: Explaining robot actions. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 187–188 (2012)

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*(3), 232–257

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332

McClure, J. (2002). Goal-based explanations of actions and outcomes. *European review of social psychology*, *12*(1), 201–235

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. Human Factors, 46(2), 196-204

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38

Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum or: how i learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547

Mitchell, T. (1997) *Machine Learning,* New York: McGraw-Hill

Morar, N. and Baber, C. (2017) Joint human-automation decision making in road traffic management, *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting,* Santa Monica, CA: HFES, 385–389

Neerincx, M.A., van der Waa, J., Kaptein, F., van Diggelen, J.: Using perceptual and cognitive explanations for enhanced human-agent team performance. In: Harris, D. (ed.) EPCE 2018. LNCS (LNAI), vol. 10906, pp. 204–214. Springer, Cham (2018)

Nunes, I. and Jannach, D. (2017) A systematic review and taxonomy of explanations in decision support and recommender systems, *User Modeling and User-Adapted Interaction*, *27*, 393–444

Paula, E.L., Ladeira, M., Carvalho, R.N. and Marzagao, T., 2016, December. Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 954-960). IEEE

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329–347

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological review*, *107*(3), 460

Rapp, B., & Goldrick, M. (2004). Feedback by Any Other Name Is Still Interactivity: A Reply to Roelofs (2004)

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1141

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*

Roelofs, A. (2004a, April). Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: comment on Rapp and Goldrick (2000). In *XIX Workshop on Cognitive Neuropsychology, Jan, 2001, Bressanone, Italy; Part of this work was presented at the aforementioned conference.* (Vol. 111, No. 2, p. 561). American Psychological Association

Roelofs, A. (2004b). Comprehension-based versus production-internal feedback in planning spoken words: a Rejoinder to Rapp and Goldrick (2004)

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 673–705

Roundtree, K.A., Goodrich, M.A. and Adams, J.A. (2019) Transparency: transitioning from human-machine systems to human-swarm systems, *Journal of Cognitive Engineering and Decision Making, 13,* 171–195

Salmon, W. C. (1971). *Statistical explanation and statistical relevance* (Vol. 69). University of Pittsburgh Pre

Sanneman, L. and Shah, J.A. (2020) A Situation-Awareness based framework for design and evaluation of explainable AI, In D. Calvaresi et al. (Eds.): EXTRAAMAS 2020, LNAI 12175, pp. 94–110

Sachan, S., Yang, J.B., Xu, D.L., Benavides, D.E. and Li, Y., 2020. An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications*, *144*, Article 113100

Sheh, R., Monteath, I.: Introspectively assessing failures through explainable artificial intelligence. In: IROS Workshop on Introspective Methods for Reliable Autonomy (2017)

Sheridan, T.B. (1992). Telerobotics, automation, and human supervisory control. *MIT press*

Sobel, J. (2020). Lying and deception in games. *Journal of Political Economy*, *128*(3), 907–947

Spiegelhalter, D., 2020, Should we trust algorithms, *Harvard Data Science Review, 2*

Sreedharan, S., Srivastava, S., Kambhampati, S.: Hierarchical expertise level modeling for user specific contrastive explanations. In: IJCAI, pp. 4829–4836 (2018)

Starke, S.D. and Baber, C. (2018) The effect of four user interface concepts on visual scan pattern similarity and information foraging in a complex decision making task, *Applied Ergonomics, 70,* 6–17

Starke, S.D. and Baber, C. (2020) The effect of known decision support reliability on outcome quality and visual information foraging in joint decision making, *Applied Ergonomics, 86*

Swartout, W. R. (1983). XPLAIN: A system for creating and explaining expert consulting programs. Artificial intelligence, 21(3), 285–325

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293

van Melle, W., Shortliffe, E. H., & Buchanan, B. G. (1984). EMYCIN: A knowledge engineer's tool for constructing rule-based expert systems. Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project, 302–313

Vasilyeva, N., Wilkenfeld, D. A., & Lombrozo, T. (2015). Goals Affect the Perceived Quality of Explanations. In *CogSci*

Vogogias, A., Kennedy, J., Archambault, D., Bach, B., Smith, V.A. and Currant, H., 2018. BayesPiles: Visualisation Support for Bayesian Network Structure Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(1), pp.1–23

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–15)

Wilson, D., & Sperber, D. (2002). Relevance theory

Windschitl, M., Thompson, J., & Braaten, M. (2008a). How novice science teachers appropriate epistemic discourses around model-based inquiry for use in classrooms. *Cognition and Instruction*, *26*(3), 310–378

Windschitl, M., Thompson, J., & Braaten, M. (2008b). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science education*, *92*(5), 941–967

APPENDIX A: NOTIONS OF EXPLANATION FROM THE HUMANITIES AND HUMAN SCIENCES

If we are concerned with the role of explanation in human engagement with AI / ML, we ought to begin our analysis with a review of disciplines that focus on this. Different disciplines take different stances in their approach to how explanation occurs, and the specific processes involved. To create a framework of explanation, we have conducted a thematic review of literature from philosophy, linguistics, social psychology, and education. By considering the different questions each of the disciplines ask, we can create a framework of explanation that is applicable to many different contexts. This builds on the work on Miller (2019) who, from a review of social science literature, concluded that:

1. people prefer contrastive explanations, i.e. a focus on choice of action rather than account of algorithm
2. people focus on a small set of features to create causal models rather than produce complete accounts with all possible causes
3. people focus on events as linear causes of outcomes rather than in terms of probabilities
4. people present explanations as part of conversations, with an expectation that they will be required to provide more information if required, rather than as unquestionable statements of fact.

A.1 PHILOSOPHY

In philosophy, to speak of explanation is to understand the nature of causality. From this, an explanation is an account in which an outcome is described in terms of its antecedent cause(s). From this perspective, there are many theories to define explanation, ranging from natural or covering laws (Hempel & Oppenheim, 1948), to statistical dependence theory (Salmon, 1971), to regularity theory (Hume, 2000), to transference theories (Dowe, 1992; Fair, 1979; Aronson, 1971). For these models, causality requires unequivocal root causes (Miller, 2019; Lombrozo, 2007; Halpern & Pearl, 2005b; McClure, 2002; Hilton, 1996).

There is, therefore, a need to fully constrain the causes before an explanation can be accepted. For example, in Hempel's covering law an explananda (i.e. an event or phenomenon to be explained) is reflected by explanans (i.e. laws that pertain to the event or phenomena and situations that apply to the specific instance of the explananda). Bird (1999) offers several examples of this idea. For example, the explananda was the fact that Mr. Smith died. The circumstances were that 'Mr. Smith ate a pound of arsenic' and the covering law is that 'Everyone who eats a pound of arsenic dies within 24 hours.' Given that the covering law applies to the circumstances, we can offer an explanation of Mr. Smith's death.

This seems to be an elegant approach, but it requires full knowledge of the circumstances and agreement on the relevance of the explananda (covering law). So, if, in a separate tragic case, Mr. Jones had eaten a pound of arsenic but had then been hit by a bus, we cannot unequivocally apply the covering law concerning arsenic (because there is uncertainty on whether the arsenic covering law or a different law concerning impact of road vehicles applies).

This suggests that there will be situations in which it would not be possible to define an unequivocal covering law. From this, and related reasons, Livengood and Sytsma (2020) argue that such definitions of

causality will not satisfy the causal attributions made in everyday explanations. So, an answer to the question 'Why is the sky blue?' would need to include a cause, or set of causes, that would describe the outcome, e.g. the sky is blue because sunlight is refracted by dust particles in the atmosphere.

However, while this might highlight some of the features that could be relevant to an explanation, it might not exhaustively detail the relationship between these features. Such a relationship needs to be inferred by the person receiving the explanation (the explainee) which means that the information that is being offered is not so much an explanation as the foundation on which to reason towards an explanation (assuming the explainee has sufficient understanding of the features and their relations in the first place).

From this perspective, the explanation is not, in effect, an answer to the question 'Why is the sky blue?' so much as a set of features that could lead one to an appreciation of an answer (providing that the explainee was able to understand the association between these concepts).

Lewis (1974) proposes that theories of explanation from philosophy agree, at least implicitly, that counterfactuals are an important factor of causality. Halpern and Pearl (2005a) extend this notion by suggesting contingent dependency. In other words, while effects may not always depend on their causes in all situations, they do depend on them under certain contingencies.

In our 'blue sky' example, the contingencies would include that the outcome is expected to occur in daytime (and not night-time), that it is expected to occur after the sun has risen and before the sun is setting, that the visibility of the sky is unimpeded by clouds, smoke etc., and that the refracted light splits into a range of wavelengths that differ in their intensity (and ability to pass through media) which produce colours visible to the human eye, etc. This improves the explanation because the counterfactuals allow the situation, in which the explanation is being provided, to be made specific. At one extreme, the issue becomes a matter of the precision with which covering laws and features of the situation are applied to produce unambiguous explanations.

For the purposes of this report, we take the idea that an explanation calls upon a set of features which define the situation, and that the relevance of these features (to an explanation) will be defined by a covering law (or, rather, some agreed principle that can be used to justify the choice of features as relevant). The use of contingencies and counterfactuals implies that the situation in which the explanation can apply will be constrained, possibly through a process by which explainer and explainee reach a consensus on the relevance of features. This accords with point 2 from Miller's (2019) list (above), but, we feel, provides more detail on the process by which this occurs.

A.2 LINGUISTICS

From linguistics, we take the stance that explanation is interactive and occurs in a conversational context. Most of the evidence for this comes from analysis of real-life conversations by Clark (1996). For Clark and his colleagues, conversations progress through the definition and maintenance of common ground which relates to mutual knowledge, beliefs, and assumptions shared by speaker and listener.

A point of contention in this definition arises from the interpretation of the word mutual; often this is taken to mean identical, but this is not at all what Clark meant. For Clark, mutual information primarily relates to having sufficient overlap (in speaker and listener's knowledge, beliefs, and assumptions) to allow a conversation to progress.

When it becomes clear that this overlap is no longer sufficient (e.g. through the use of back-channelling by the addressee to indicate disagreement or lack of understanding), then the conversation needs to re-establish common ground (e.g. through the speaker

either introducing additional information or clarifying current information). Implicit in this notion of common ground is the idea that speaker and addressee will apply just enough cognitive effort to maintain the flow of the conversation.

This is similar to Grice's (1975) cooperative principle, which describes how people achieve effective conversational communication in common social situations. For Grice (1975), conversations appear to progress in terms of four maxims, namely: Quality (truthfulness), Quantity (informativeness), Relation (relevance) and Manner (clarity). Speakers are expected to observe and adhere to these principles (Davies, 2007; Wilson & Sperber, 2002; Grice, 1975) to make a contribution to the conversation that is accepted as relevant.

Sperber and Wilson (2002) argue for a cognitive principle of relevance which states that human cognition is evolutionarily geared towards the maximisation of relevance. According to Wilson and Sperber (2002), an input is relevant only when its processing yields positive cognitive effects; so, when an input connects to background information available to individuals, they can produce conclusions that are relevant to them. Furthermore, an input is worth picking out only when it is more relevant than any other input available to the individual at the time (the greater the positive cognitive effects, the greater the relevance). Relevance also depends on the effort required to process the input – the more effortful the processing of an input, the lesser its relevance. The challenge, from this perspective, lies in defining the processes by which relevance is determined (beyond the suggestion that individuals can attend to information in terms of cognitive effort). This accords with point 4 from Miller's (2019) list.

Critics note that Sperber and Wilson's (2002) definition of relevance is too vague to be applicable in natural setting and, thus, leaves the concept too open to interpretation to be useful. One solution to this might be to draw a parallel with Hempel's covering law . From this, relevance could be considered in terms of a) the circumstances in which an explanation is being presented (i.e. those features of the circumstance to which explainee is attending), and b) the law by which these features are deemed to be relevant.

While this produces a clumsy, circular argument (to the effect that features are relevant because they are relevant because of the law that relates them to the circumstances) and does not offer much to help with understanding the cognitive processes, it does offer support for the role of bidirectional conversation between explainer and explainee. To illustrate this, the explainee asks 'Why is the sky blue?' and the explainer speaks of refraction and dust particles. If the explainee does not know the meaning of the word refraction then this creates a cognitive demand (perhaps to use their knowledge of physics of light to guess what it means) or an indication to the explainer that there is a problem (meaning that the explanation in its current form does not satisfy Grice's cooperative principles) and, thus, requires more information.

From this we assume that an explanation will involve clarification of features (reinforcing the notion of contingency dependence from philosophy) and consensus on these and the definition of relevance. If consensus as to whether information is relevant arises from the conversation between explainer and explainee, then this might involve more than 'laws' because there might be agreement as to what constitutes relevance (and this agreement might involve imprecise or informal definition of the relationship between features and circumstance).

Therefore, we propose (as a starting point) that relevance could begin with seeking agreement on specific features that the explainer and explainee agree would be appropriate, and, if there are several features, then they might be grouped, which we will call cluster, or by an expectation (based on prior experience) that the features occur together, which we will call belief, or by a covering law by which features must occur together (which we will call policy) and which can allow action. From this perspective, a cluster exists purely in terms

of the co-occurrence of features and has no capability of making predictions or drawing inferences about the features. Most of the simpler ML algorithms rely of structuring features in terms of regression, correlation, or similarity. A belief, on the other hand, considers the co-occurrence of features in terms of prior experience and can be used to make predictions about what might happen if specific features were altered. Therefore, clusters should be considered separate to beliefs.

A.3 SOCIAL PSYCHOLOGY

Social psychology assumes that, like linguistics (*Section A.2*), explanation occurs in a social context (Miller, 2019; Malle, 2006; Hilton, 1990). People are more likely to offer one or two features as first-pass explanation (McClure et al., 2001; McClure & Hilton, 1997; Leddo, Abelson, & Gross, 1984; Tversky & Kahneman, 1983). These features imply a) a string of causal reasoning that the other people are assumed to be able to perform and b) are sufficient to explain the situation. The former relates to the notion of common ground and the latter relates to the notion of relevance. Thus, both the explainer and explainee are actively involved in the explanation process.

Further support for the idea of a string of causal reasoning can be found in Sobel's (2020) research on deception. In this work, Sobel notes that to lie only requires the existence of accepted message meanings, to deceive requires a model of how the audience will respond to the message, and to cause damage requires an appreciation of the consequences that will ensue in comparison to providing better information. This emphasises the focus on outcome or action as per point 1 of Miller's (2019) list.

Research into how people draw causal connections echoes the discussion of causality. The key difference here being that social psychologists are interested in how humans make causal connections as opposed to philosophy's concern with the fundamental nature of causality. In social psychology, the focus of the explanation is the behaviour of other people, e.g. why did person X do act Y? Malle's (2006) theory of social explanation argues that humans make social attributions: people attribute behaviour of others and themselves by assigning specific emotional or mental states to the person performing that behaviour, e.g. individual X shouts at individual Y. The behaviour (shouting) could, perhaps, occur because individual X is angry or individual Y is in danger.

To select a state to attribute to individual X, we might need to know the circumstances in which the behaviour occurs, and we might also draw on prior knowledge of behaviour performed in similar circumstances or of behaviours performed by individual X. Malle (2006) argues that if the explainer believes person X's action was intentional, they will ascribe motivations to the individual. For example, individual X shouted to warn individual Y of an oncoming bus. If the behaviour was believed to be unintentional, the explainer will offer just causes, such as physical, mechanistic, or habitual causes.

For example, individual X shouted as the result of being hungry and annoyed that the computer was not formatting a document properly and being interrupted by individual Y. This suggests that the relevance (borrowing from the discussion of linguistics in *Section A.2*) can vary according to the situation.

Keil (2006) notes that qualitatively different patterns of explanation can be used in talking about domains such as physical mechanics, biological function, or social interactions. Therefore, it is important to recognise that the way causal connections are drawn may alter depending on the situation (and knowledge of the explainer and explainee).

Causal model theory (Rehder, 2003) suggests causal connection is the explicit representation of the probabilistic causal mechanisms that link category features and objects by evaluating whether they were likely to have been generated by those mechanisms. The underlying assumption of causal model theory is that individuals use abductive reasoning to infer

explanations. Chin-Parker and Cantelon (2017) show that counterfactual reasoning could also be involved in causal connections involving categorisation and their research suggests that counterfactual reasoning could be key when drawing initial causal connections.

Depending on the situation there will be constraints by which information for the explanation is selected. Such constraints can be categorised as either implicit (e.g. cognitive effort in constructing an explanation, explainer goals, and use of contrastive examples) or explicit (e.g. access to information). Therefore, it is likely that the selection criteria used in each individual case of explanation is dependent on the situation, the goals of both the explainer and explainee, the context of the problem, and various other subjective precursors. In this respect, relevance will be influenced not only by the most appropriate causal model to apply to the circumstances but also by the most appropriate model for the abilities of the explainee.

This relates to the discussion of linguistics and adds a further factor concerning the perceived knowledge and ability of the explainee. If the explainer assumes too little knowledge, then the explanation could appear patronising; if the explainer assumes too much knowledge, then the explanation could be opaque. We consider how explainers adapt their explanation to explainee knowledge in our discussion of education (*Section A.4*).

Hoffman and Klein (2017) suggested that explanations can come in different forms (sentence, list, narrative, diagram etc.) and in varying levels of depth (local vs. global). How an explanation is conveyed is then dependent on the situation, the explainee, and the question asked accordingly. However, we believe this should be taken a step further and, therefore, suggest incorporating this idea with relevance. So, in terms of explanation, the explainer will choose the features deemed most relevant to an event, how to present the information and the level of explanation based on the ability to connect with the context of the explanation availability and how effortful they would be to produce for explanation.

Explanation evaluation is typically assumed to occur from the perspective of the explainee, although, given the bidirectional conversation that surrounds explanation, one might assume evaluation will also be performed by the explainer. So, while one could argue that some of the criteria in explanation selection are valid for explanation evaluation, differing goals of the explainer and explainee could impact the criteria used to evaluate the explanation.

Therefore, it is important to acknowledge that there are some differing processes occurring within the explainer and explainee during an explanation. For example, Hilton (1996) and Jaspars and Hilton (1988) argue that a good explanation must be relevant to the question asked and the mental model of the explainee, while Vasilyeva, Wilkenfold, and Lombrozo (2015) say that the goals of the explainee also affect the criteria for evaluation. Miller (2019) concludes that the most agreed upon and important criteria for explanation evaluation are probability, simplicity, generalisability, and coherence with prior beliefs. Note, however, that probability is only important to an extent.

For clarity on this, take the comparison Miller (2019) makes: a student coming to their teacher to ask why they received 50% on an exam. An explanation that most students scored around 50% is not going to satisfy the student. Adding a cause for why most students only scored 50% would be an improvement. Explaining to the student why they specifically received 50% is even better, as it explains the cause of the instance itself. So, the question is whether to explain events using generalisations, or to use specific instances.

It is also worth noting that while a true / likely cause is an attribute of a good explanation, to say that the most probable cause is the best explanation would be incorrect (Hilton, 1996). Social psychology, unlike philosophy, seems to not only acknowledge that people will draw causal connections that do not necessarily

reflect true causation, but that there are different strategies for doing so. Therefore, a good framework for explanation should allow for flexibility in causal connectivity as well as accounting for the fact that humans can imply causation incorrectly. Similarly, in explanation selection, a good framework should be flexible enough to accommodate the changes in the selection criteria dependent on the precursor information.

Another important finding from social psychology is that the explainee does not use the same criteria to judge an explanation as the explainer uses to select their explanation. Therefore, this further supports the idea that the explainee should be considered as an active part of the explanation process.

A.4 EDUCATION

As one may presume, the main question about explanation asked in the education literature by researchers is 'What is the most effective way of explaining?' While philosophers argue over which theory of causality is correct, researchers in education have found that most of the major theories of causality are useful for aiding student understanding in different ways (*Table 6*).

Education researchers distinguish between practices focused on helping students' understanding of explanations from authoritative sources (such as texts and teachers), and dialogic practises focused on having students create scientific explanations based on their own ideas and understanding of evidence (Windschitl, Thompson, & Braaten, 2008a; Windschitl, Thompson, & Braaten, 2008b). This highlights distinctions between the purpose of the explanation and the differing roles of stakeholders in the explanation process.

The complexity of information is also important for fostering understanding in students. Vygotsky's (1980) work suggests that students learn best when the knowledge presented to them is within their zone of proximal development (ZPD). Essentially, the ZPD is knowledge the student can acquire and understand with help from a more knowledgeable other. Using scaffolding techniques, such as engaging the student by questioning them and by providing analogical reasoning examples of a process, can aid students with the acquisition of knowledge (Kelemen, 2019;

Type of explanation	Pros and Cons
Covering Law	Can allow teachers to gain insight into student understanding and help them further their attempt at explanation.
	Does not aid conceptual reasoning or theory building abilities. Additionally, it does not accommodate explanations of unlikely events.
Probabilistic model	Engages students in important data analysis practises and allows teachers to engage students in data interpretation and scaffold their attempts at inferring from data.
	However, the emphasis on statistics may not acknowledge underlying events.
Causal model	Engages students in theorizing about unobservable causes for observable phenomena and allows teachers to engage students in model building.
	There is a tendency toward developing only simple, linear cause–effect relationships instead of causal webs and models.
Pragmatic approach	Encourages explicit communication between teachers and students about locally constructed norms and meanings for explanations affording sense-making opportunities.
	It does not subscribe to a theory of explanation – no theoretical backing.
Explanatory unification	Useful way of getting students to focus on "big picture" ideas.
	Requires the explainee to have sufficient necessary background information to reason and explain in this way, so is only for more advanced learners.

Table 6. The advantages and disadvantages of using different theories of causality for explanation. Adapted from Braaten and Windschitl (2011).

Arias, Davis, Marino, Kademian, and Palincsar, 2016). Scaffolding techniques and Vygotsky's theory are widely used in education today and the ideas have received a lot of empirical support (Windschitl, Thompson, & Braaten, 2008a; Windschitl, Thompson, & Braaten, 2008b).

The process of scaffolding, in terms of defining the knowledge held by the explainee so the explainer can provide information that can be assimilated and accommodated, is reminiscent of Clark's notion of common ground. Evidence from the education literature suggests that a good framework of explanation should account for how the goal of the explainer can account for changes in the explanation structure.

Furthermore, education promotes the idea that a good explanation should consider the knowledge of the explainee. If the goal of the explainer is to expand the knowledge and understanding of the explainee, techniques to aid this should be accounted for in a good framework of explanation.

A.5 HUMAN FACTORS

In human factors, researchers have drawn an analogy between understanding the workings of ML and situation awareness (Endsley, 1995). In broad terms, situation awareness means gathering information from a specific situation (e.g. an aeroplane cockpit) and using this information to define the immediate events (e.g. the state of the aircraft) through appreciation of how the situation has arisen and prediction of how this situation might change in future. This, in Endsley's theory (1995), is defined in terms of levels:

1. Perception of current situation
2. Comprehension of how the current situation evolved to this state
3. Projection of changes

Thus, Chen et al. (2014) define situation awareness transparency as "…the quality of an interface to support a human operator's comprehension of an intelligent agent's intent, performance, future plans and reasoning process." [p. 2]. Similarly, Sanneman and Shah (2020) and Roundtree et al. (2019) define transparency in terms of levels of situation awareness:

- Level 1 – In this level, situation awareness (of the human operator) focuses on the current state of the AI system. In part, this involves addressing 'what' questions, e.g. defining what the AI system did or is doing in terms of the data that it is using or the results that it has produced. This is what Roundtree et al. (2019) terms 'seeing through the system' and involves "…sensitivities to inputs, semantic feature information or model representations, cluster information, or abstracted representations of model details." [Miller et al., 2017, p. 99]
- Level 2 – In this level, the human operator seeks to comprehend the rationale for the AI systems results. This addresses 'why' questions and involves what Roundtree et al. (2019) termed 'seeing into the system'. This could involve, for example, an appreciation of the underlying algorithms and their settings through the exploration of interim results, as well as an appreciation of the context in which the algorithms have been applied (e.g. in terms of the timeliness and coverage of the data, and the limits and constraints of the algorithms, etc.)
- Level 3 – In this level, the focus is 'what if' and 'how' questions. This addresses predictions of what the AI system might do next or how the AI system might respond to changes in data or situation, and how it might respond if the human operators do, or do not, follow the recommendation. This could involve contrastive or counterfactual cases to explore 'what if' scenarios for the AI system.

In addition to a focus on explanation and situation awareness, other human factors approaches draw on the consideration of human interaction with automation (*Section 3*) and consider stages of explanation (Anjomshoae et al., 2019; Neerincx et al., 2018), types of errors (Sheh and Monteath, 2017), agent cognitive

states (Harbers et al., 2012), and the theory of mind (Hellström and Bensch, 2018).

A.6 THE DATA / FRAME MODEL OF SENSE-MAKING

Central to sense-making in the data / frame model (illustrated in *Figure 13*) is the relationship between the data to which the analyst has access and the different frames that can be used to interpret, make sense of, or explain these data (Klein et al., 2006a).

A key stage in sense-making involves deriving a sufficient understanding of the situation to be able to match it to an appropriate schema. In the data / frame model, a frame is applied to a set of the data, or a set of the data could suggest a frame. This reciprocity points to the continuous interweaving of exploring data and generating interpretations. Kang and Stasko (2011) note that "…analysis is about determining how to answer a question, what to research, what to collect, and what criteria to use." [p. 25]. The point at issue is not how people answer questions but how they define them in the first place (Roth et al., 2010). The Intelligence (or Analysis) Cycle (NATO, 2008) involves four phases:

1. Direction – definition of objectives for gathering intelligence through intelligence requirements and requests for information.
2. Collection – gathering and receipt of information by agents in response to the intelligence requirements or through more spontaneous and serendipitous routes.
3. Processing – compiling and interpreting information to produce intelligence.
4. Dissemination – distribution of appropriate parts of the intelligence to relevant parties.

While this intelligence cycle might begin with direction, this only gives a high-level sense of what the analyst might be looking for. As collection and processing progress, new opportunities arise through discovery-led refinement (Attfield and Blandford, 2010). Heuer

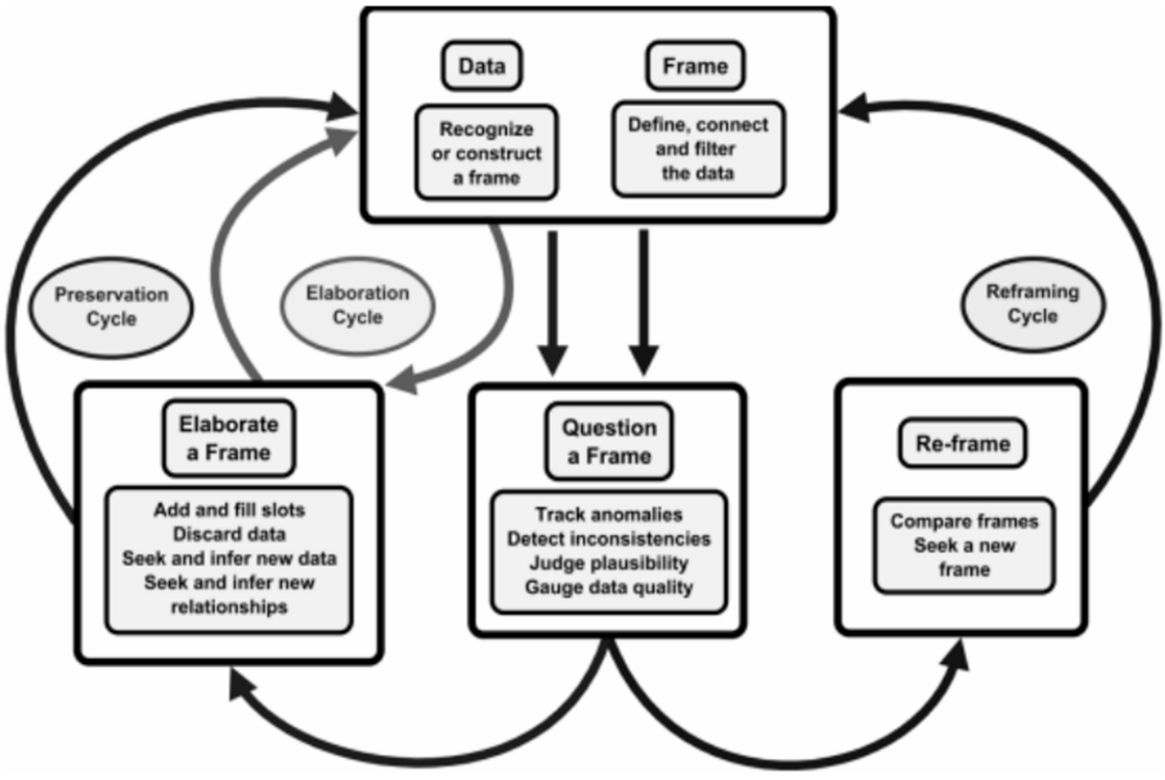


Figure 13. Data / frame model of sense-making

(1999) distinguishes between data driven analysis (i.e. applying well-understood analytic procedures to well-defined data sets) and conceptually driven analysis (i.e. dealing with complex, ambiguous, and uncertain data).

Conceptually driven analysis implies a cycle of activity that involves "the reciprocal interaction of information seeking, meaning ascription and action" [Thomas, Clark and Gioia, 1993, p. 240]. Elm et al. (2005) define this activity in terms of down-collect (sample from the available data for material deemed to be on analysis), conflict and corroboration (ensure accurate and robust interpretation of findings and modify the down-collect accordingly), and hypothesis exploration (construct coherent narrative to explain the findings and reflect this narrative back to the conflict and corroboration activity).

Kang and Stasko (2011) identified four main activities (noting that these activities overlapped and intertwined as the project developed):

1. Construct conceptual model of issues
2. Collect information
3. Analysis
4. Report findings

Similarly, Baber et al. (2018) demonstrate that the experienced intelligence analysts will continually test their interpretation of their analysis through practice briefings, working through an alternating sequence of broad high-level questions and narrow feature-specific analyses to refine the conceptual model of the analysis.

APPENDIX B: NOTIONS OF EXPLANATION IN MACHINE LEARNING

The challenge of enabling people to understand the output of computer algorithms has gone hand in hand with the development of computing technology. Once computers could, in some sense, work autonomously on data then there was a need to appreciate not only the solution that the computer had produced but also the rationale for this solution (if only to determine whether any disagreement between the actual and expected results were due to problems with the algorithm or with the data on which the algorithm had been applied). Thus, there is discussion of explanation in expert systems (Swartout, 1983; van Melle et al., 1984), context-aware computing (Bellotti and Edwards, 2001), recommender systems (Cramer et al., 2008; Herlocker et al., 2000) and a call for Explanation-aware Computing (EXaCT 2008). As a starting point for the ways in which the contemporary ML community approaches explanation, it is worth considering how Langley et al. (2017) consider the operation of an agent that can produce an explanation:

1. Given a complex set of objectives that require an agent's extended activity over time
2. And given background knowledge about categories, relations, and activities that are relevant to these objectives
3. Produce records of decisions made during plan generation, execution, and monitoring in pursuit of these objectives
4. And produce summary reports, in human accessible terms, of the agent's mental and physical activities
5. And produce understandable answers to questions that are posed about specific choices and reasons for them.

To perform these steps, the 'explainable agent' needs to be able to represent its knowledge in a way that supports explanation so that a human can understand, needs to represent changes in this knowledge in episodic memory, and needs to be able to use this episodic memory to answer questions that require rationale for actions.

Some ML algorithms rely on assumptions about the data on which they are trained. For example, there might be an assumption on the likelihood of features that are being clustered or how closely the features would need to be related to fit a cluster. Some ML algorithms allow users to modify the number of clusters to produce, or the degree of closeness between features. In both cases, when the algorithm produces an output, it does so based on the assumptions and settings.

The person who uses the output might not be aware of these assumptions or settings or might not realise how changing these affects the output. Furthermore, some data sets might involve very sparse instances of some of the features. For example, there might be many instances of normal events and few instances of unusual events. In these circumstances, it is common to use sampling techniques (Hacibeyoglu and Ibrahim, 2018) which can involve generating artificial entries for the minority class (over-sampling), removing entries from the majority class (under-sampling), or a hybrid approach combining the two. While this can resolve problems that the data cause for the algorithm, it can result in a data set that deviates from the original set (such that, for instance, if one was to produce summary statistics on the over- or under-sampled set the results would differ from the original – in the same way that transformation of the data (e.g. to allow assumptions of normality) will alter the summary statistics). Unless the sampling or transformation approaches are made clear to the consumers of the algorithm's output, there

might be misunderstanding or confusion of the features that are being used.

Sridharan and Meadows (Sridharan and Meadows 2019) state the following principles for explanation generation:

1. Explanations should present context-specific information relevant to the task, domain, and the question under consideration at an appropriate level of abstraction.
2. Explanations should be able to describe knowledge, beliefs, actions, goals, decisions, rationale for decisions, and underlying strategies or models in real-time.
3. Explanation generation systems should have minimal task or domain-specific components.
4. Explanation generation systems should model and use human understanding and feedback to inform its choices while constructing explanations.
5. Explanation generation systems should use knowledge elements that support non-monotonic revision based on immediate or delayed observations obtained from active exploration or reactive action execution.

The interpretation of explanation (and its associated concepts as outlined in *Table 1* and *Table 2*) depend on the nature of the ML algorithms that are being applied; some of these might rely on correlations within data sets, others might extend the notion of correlation to seek to define clusters in these sets, and others might seek to maximise reward functions. From this perspective, a focus of explanation is on either the data set being explored or the operation of the algorithms themselves. So, for example, Gunning and Aha (2019) define explainable AI (XAI) in terms of "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future."

To formalise the relationships within XAI, Holzinger et al. (2020) propose a process model (*Figure 14*). In this the human, h, or machine, m, produces a statement, s, such that $s = f(r, k, c)$ – where r: representation of unknown fact relating to an entity, U_e; k: pre-existing knowledge; c: context.

In this model, the ideal state is when $m_h = m_m$ = ground truth (gt) = $s_h = s_m$. However, gt is seldom fully defined, and the models that humans create *tend* to be causal and the models that machines create *tend* to be relational (e.g. correlation, regression, distance, similarity). From this, explainability "…highlights decision relevant parts of machine representations, r_m, and machine models, m_m – i.e. part which contributed to model accuracy in training or to a specific prediction. It does not refer to a human model, m_h." (Holzinger et al.,

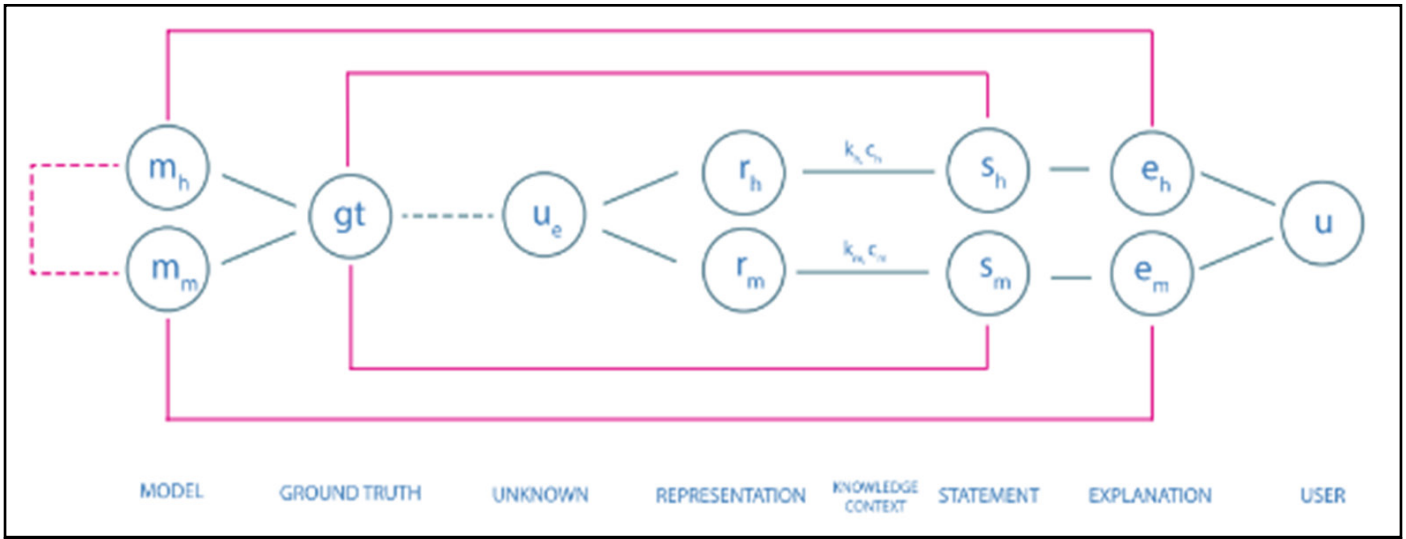


Figure 14. A process model of explanation [Holzinger et al., 2020]

2020, p.195). Leaving aside the omission of the human model (which we revisit in *Section 6*), an interesting aspect of this process model is its recognition of the ways in which aspects of the situation (defined by a ground truth) are represented (by human or machine) to create a statement that provides an explanation. This is, in principle, like the covering law and situation awareness; possibly a combination of these concepts, something akin to the data-frame model in (Section 4).

Our review argues (in *Section 4*) that explanations offered by humans will rarely, if ever, begin from the perspective of demonstrating the process by which a conclusion is reached. That is, human explanation seldom begins with the algorithm (even in education), but, rather, begins with the data that support the algorithm and with instances that enable the limits of the algorithm to be explored (i.e. counter-factual cases). As the discussion of linguistics suggested, human explanation tends to seek minimal cognitive burden (on explainer and explainee) and, by implication, XAI also has an overarching goal of transparency without overhead to increase the trust that humans will place in the AI (Borgo et al., 2018; Fox et al., 2017; Lipton, 2016). From this, approaches which focus on evaluation of algorithm outputs side-step the question of explanation (or, at least, imply that an explanation is equivalent to a repeatable output of the algorithm). This implies a series of trade-offs, e.g. between transparency (in terms of algorithmic complexity), explainability, and trust.

One way of conceptualising this is in terms of performance vs. interpretability, as shown in *Figure 15*. In some versions of this figure, the phrase performance is used in preference to algorithmic complexity, but this begs the question of whether more complex algorithms always result in better performance. What we can note is that the algorithms with higher complexity (which either involve black box, such as deep learning, reinforcement learning, recurrent or convolutional neural networks, or involve combinations of several algorithms, e.g. random forests with their use of many decision trees) are difficult for humans to simulate. While the algorithms that make use of regression or correlations can (to some extent) be simulated by human observers (in that the data that are used and the mathematics applied to these data can be appreciated

Accuracy	How well an explanation predicts unseen data
Fidelity	The explanation ought to be close to the predictions of the explained model
Consistency	The explanation should apply equally well to any model trained on the same data set
Stability	When providing explanations to instances, similar instances should produce similar explanations
Representativeness	A highly representative explanation is one that can be applied to several decisions on several instances
Certainty	If the model at study provides a measure of confidence on its decisions, an explanation of this decision should reflect this
Novelty	This property refers to the capability of the explanation mechanism to cover instances far from the training domain
Salience (of features)	The explanation should pinpoint the important features

Table 7. Aspects of evaluating an algorithm

	Label	Description
User Preferences and Input	Decisive Input Values	Indication of the inputs that determined the resulting advice.
	Preference Match	Provision of information about which of the user preferences and constraints are fulfilled by the suggested alternative.
	Feature Importance Analysis	Justification of the advice in terms of the relative importance of features, e.g. by showing that changing feature weights would cause the selected alternative to be different.
	Suitability Estimate	Indication of how the system believes that the user would evaluate the suggested alternative, e.g. by showing a predicted rating.
Decision Inference Process	Inference Trace	Provision of details of the reasoning steps that led to the suggested alternative, e.g. a chain of triggered inference rules.
	Inference and Domain Knowledge	Provision of information about the decision domain or process, e.g. about the main logic of the inference algorithm. For example: "We suggest this option because similar users liked it."
	Decision Method Side-outcomes	Provision of algorithm-specific outcomes of the internal inference process, e.g. a calculated number that expresses the system's confidence.
	Self-reflective Statistics	Provision of facts regarding the system's performance, e.g. by informing the user how many times the system made decision suggestions in the past that were accepted.
Background and Complementary Information	Knowledge about Peers	Provision of information about the preferences of related users, e.g. ratings given to a suggested alternative by social friends.
	Knowledge about Similar Alternatives	Indication of similar alternatives that were an appropriate (system's or user's) decision in a similar context in the past, e.g. items that the user or related peers showed interest in.
	Relationship between Knowledge Objects	Provision of information about the relationship between features, or features and users. This can be done, for example, in the form of a directed acyclic graph representing a causal network.
	Background Data	Provision of (external) background data specific to the current problem instance, e.g. data derived from processing posts in a social network, which were considered in the decision inference process.
	Knowledge about the Community	Provision of information that supports the decision based on the behaviour and preferences of a community, e.g. showing the general popularity of the proposed alternative.
Alternatives and their Features	Decisive Features	Indication of the features of the alternative that are key to the decision.
	Pros and Cons	Indication of the key positive and negative features of the alternative.
	Feature-based Domination	Justification of a decision in terms of the dominance relationship between two alternatives, e.g. by showing that an alternative is not selected because it is dominated by another.
	Irrelevant Features	Indication of features that are irrelevant for the decision, typically when the values of such features in the suggested alternative are not considered good.

Table 8. Types of information used in explanations in recommender and decision support systems [Nunes and Jannach, 2017]

sufficiently to be able to either imagine the result or calculate instances by hand), the algorithms which maximise reward tend to be more opaque (and can be referred to as black box models).

EXPLANATION AND INTERPRETABILITY

While we used interpretability in *Figure 15*, it is important to note that this is *not* a monolithic concept, as Lipton (2016) points out, but consists of different levels of transparency. At the level of the entire model, the ability of the human to simulate the operation of the entire model contributes to the model transparency. It is worth noting, at this point, the simulatability (in this context) becomes a function of the model's transparency *and* the knowledge, skills, and abilities of the specific human, e.g. in terms of how well the person understands the situation, or the data, or the algorithm (and different stakeholders may vary in

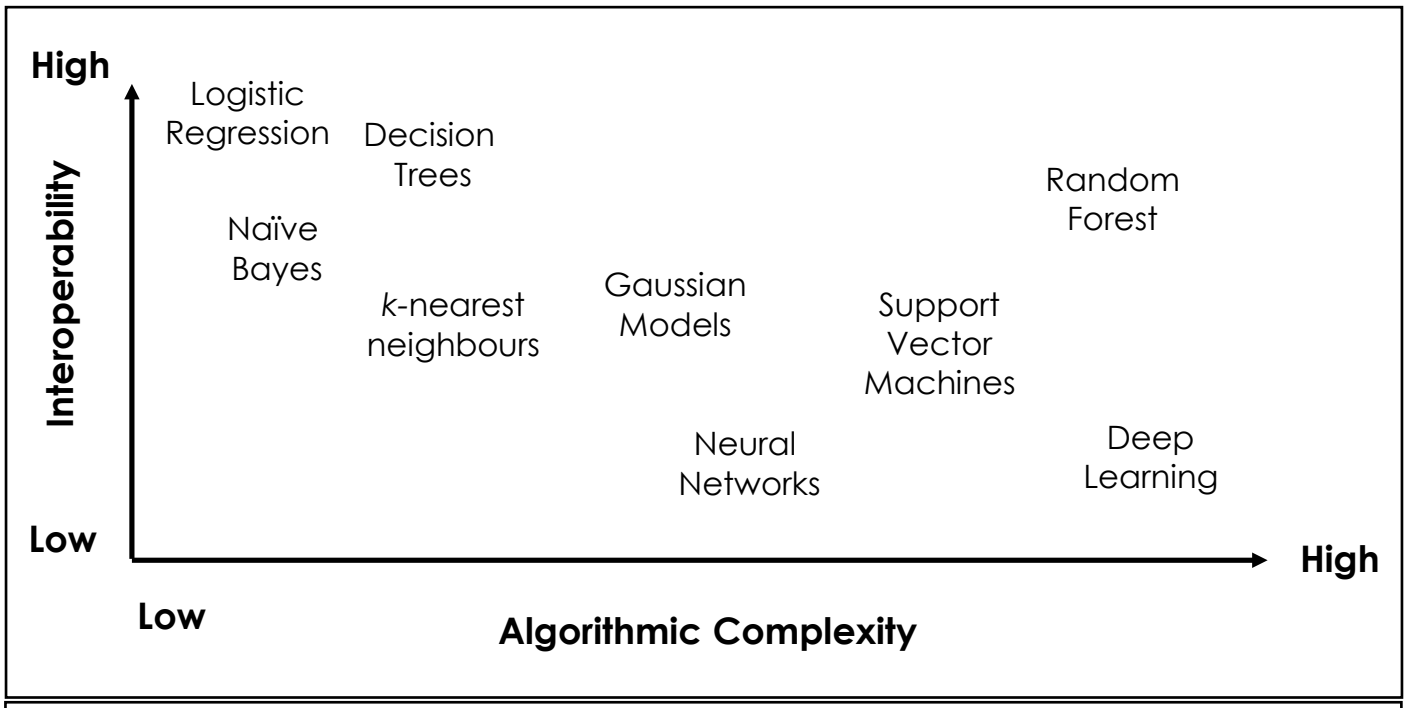


Figure 15. Contrasting ML algorithms by interpretability and algorithmic complexity

terms of their level of understanding for each of these aspects). At a lower level, each part of the model might have their own transparency. In part, this creates local simulatability (for each part of the model) but also relates to the decomposability of the whole model to its constituent parts. In this case, transparency might relate to how easily the human can apply intuitive understanding of the input data or the settings that the model applies to these.

A potential problem here might be that the model weights might be situation dependent (meaning that understanding the impact of the weights in one situation might not generalise to other situations), or these weights might not be understandable but the output (say, in the form of a regression plot) might be – which could mean that the understanding is based on the human's ability to impose meaning of the output (rather than knowledge of the model). Furthermore, the algorithms applied (either in parts or to the whole model) have their own degree of transparency.

Although black box models usually perform better than simpler (but intrinsically explainable) ML algorithms, such as linear regression, logistic regression, and decision tree (Chen et al., 2018; Ribeiro et al., 2018), people tend to trust the latter more, because they are easy to understand. That is, poor interpretability of black box models can impair willingness to deploy these potentially more accurate models. This might be particularly true in high-stakes scenarios such as self-driving cars, medical diagnosis, criminal investigation and profiling, and financial fraud (Molnar, 2019; Singh et al., 2020; Deeks, 2019) when the explanation becomes part of an auditable investigative process.

A survey of medical practitioners (Tonekaboni et al., 2019) suggested that "Clinicians overwhelmingly indicated that the model's overall accuracy was not sufficient." Clinicians want to know the reason behind the model prediction. The interpretability could certainly increase social acceptance and could be used to manage social interactions (Doshi-Velez and Kim, 2017). Even models that fall short in accuracy were deemed acceptable so long as there is clarity around why the model under-performs (Tonekaboni et al., 2019). Moreover, a single metric, such as classification accuracy, is an incomplete description of most real-world tasks. Although the typical way to evaluate the model – accuracy, precision, recall, F1-score – could estimate the quality of a model in many cases, it is worth noting that there are several ways

these evaluations can go wrong (e.g. because of bias and errors due data leakage, or data set shift arising from separation of training and validation of data set). However, people could be more sensitive to the errors and bias in if they were provided explanations and could correspondingly modify models. In order to address this urgent, interpretability tools (IT) are being developed (Bhatt et al., 2020; Singh et al., 2020; Tonekaboni et al., 2019).

From the field of robotics, researchers have sought to allow humans to ask questions of the robot that allow it to explain its planning and reasoning processes. For example, Fox et al.'s (2017) Explainable Planning (XAIP) requires the robot to justify why the planner chose specific actions over others, why it believed certain actions could not be executed, and why / if replanning is needed, etc. This approach tends to produce overly verbose statements that do not always reflect the nuances of the situation (relying, instead, on the knowledge structures that the robot employs). To address the problems of verbosity, Amir and Amir (2018) apply the HIGHLIGHTS program to generate summary reports of the robot's decision-making over time (as opposed to explanations of each individual decision). However, this means that the robot is compromised on its ability to address specific decisions.

Recognising that human explanations often rely on contrasts (between choosing action X or choosing action Y in a situation), Borgo et al. (2018) developed XAI-Plan to provide the rationale for a robot's initial plan, with options for actions which can then be used to justify the selection of one action over another in a given situation. A similar approach, developed by Krarup et al. (2019), generates a possible plan based on the human's question (i.e. by treating the content of the question as a foil or counterexample) and compares this possible plan with the one that had been followed. This use of contrastive explanations allows 'why' questions to be addressed. Korpan and Epstein (2018) developed the Why-Plan system to allow humans to ask questions of robots performing navigation tasks. While the motivation for this work seems to be to work from the assumption that human explanations involve facts and foils (i.e. evidence and counterexamples), it is not clear that the ensuing dialogue is natural. Nor is it apparent that the ability to answer 'why' questions necessarily involves generation of a foil, even if it is implicit (Hilton and Slugoski 1986, Lipton 1990, Hilton 1990, Lombrozo 2012).

Alternative approaches to the provision of explanation by robots focus on stating the purpose or goal that the robot is seeking to achieve (McClure 2002; McClure et al. 2003; Dannenhauer et al. 2018a; Dannenhauer et al. 2018b). To do this, the robot needs to be able to state beliefs (in terms of what information it has obtained

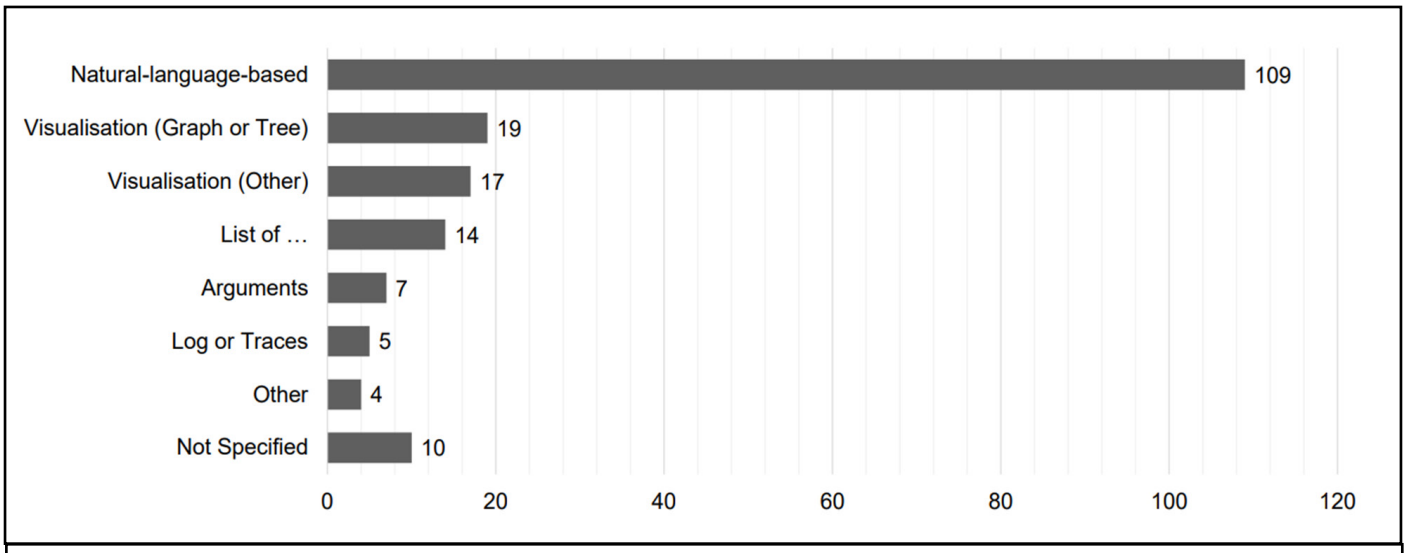


Figure 16. Primary display modes for recommender and decision support systems [Nunes and Jannach, 2017]

from the environment and how it has interpreted this), desires (in terms of the purpose or goal it is seeking to achieve), and intentions (in terms of the plan that it will apply to achieve the goal). This takes us towards the concept of explicable planning: in terms of being able to generate plans that are amenable to human explanation; and explanation generation: the ability to tailor explanations to humans with given knowledge (Chakraborti et al. 2018; Chakraborti et al. 2017; Sreedharan et al. 2017).

ML and AI communities are developing along three broad lines of explanation:

1. Approaches which are sufficiently simple to support simulatability (e.g. linear or logistic regression, decision tress, naïve Bayes, Hidden Markov Models) either because the human operator can calculate examples of these or because the models can be visualised in ways that make cause-effect relations easy to appreciate.
2. Approaches which present the goals or plans by which the agents are reasoning. These have been considered in the preceding discussion.
3. Approaches which allow induction of models, i.e. which can allow simplification of the underlying model or modification of weights in the model to reveal how features can be changed to apply different rules.

Approaches to explanation could also focus on the effect of one or two features on predicted outcome. Feature importance is used across many different domains – finance, healthcare, facial recognition, and content moderation. Also known as feature-level interpretations, feature attributions, or saliency maps, this method is by far the most widely used technique (Gilpin et al., 2019; Baehrens et al., 2010) and is highly requested by ML practitioners, like what was investigated in the medical domain "Clinicians repeatedly identified that knowing the subset of features deriving the model outcome, is crucial." (Tonekaboni et al., 2019). Approaches to identifying feature weighting include partial dependence plot (PDP) (Friedman, 1999), individual conditional expectation plot (ICE) (Goldstein et al., 2014), and accumulated local effects plot (ALE) (Apley and Zhu, 2020), etc. These are illustrated in *Figures 10, 11 and 12.*

By considering specific instances that could be analogous to other instances (either in k-nearest space, Caruna et al., 1999) or using case-based reasoning (Kim et al., 2014; Doshi-Velez et al., 2018) one can provide explanations by example. This assumes that the user can not only spot the similarity between the examples that have been defined as analogous, but also reason as to why these analogies have been drawn.

In terms of model induction, a popular approach involves some form of explanation by simplification, in which a surrogate model is built from the trained model to be explained. Explanation, from surrogate techniques, could be categorised as global or local. Global explainability attempts to understand the high-level concepts and reasoning used by a model. Local explainability aims to explain the model's behaviour for a specific input (Guidotti et al., 2018). Surrogate techniques characterise the concepts learned by the model to create simpler models and then explain the model's outcome by listing those features which are most relevant to the outcome. The ideas described above have been implemented in feature extract interpretability tools such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016). LIME builds an explanation around a specific solution instance, i.e. a local prediction, in which the probability that the instance belongs to a specific class is a function of the features around this instance. In this way, sparse linear models can be interpolated from the data set to produce a simplified version of the output (which focuses on the relations between these data around the specific instance rather than the model as a whole). For this latter reason, the approach is model-agnostic and concentrates on local fidelity, i.e. the relations within the vicinity of the specific instance. This could mean that the features selected might not always apply in other instances, i.e. that locally

important features might not be globally important. However, it does mean that the explanation can be couched in terms that the human is able to interpret and understand. In effect, the approach echoes the logic of *Figure 14* and of the covering law.

An extension of this approach is to create a set of instances which can be used to anchor subsequent reasoning, i.e. Anchor Local Interpretable Model-Agnostic Explanations (aLIME) which can be presented in the form of production rules, i.e. if X then Y. In addition to a surrogate model, we could explain model outcome by computing the contribution of each feature to the prediction from a game theory aspect. The idea is adding the feature value that would contribute the most to the prediction and iterate until all feature values are added. Shapley Additive Explanations (SHAP) creates an explanation in terms of the contribution of each feature to the prediction which then assigns each feature an importance value (Lundberg et al., 2017). Similarly, ELI5 (Fan et al., 2020) provides a way to compute feature importances for a black box estimator by measuring how score decreases when a feature is not available. Visualisations of LIME and SHAP are shown in *Figure 17.*

For black box models (which use deep learning forms of AI), it might not be possible to define specific features that the models use. Consequently, the outputs of these models can be considered in terms of actions the model performs (e.g. making a move in a board game). From this, the human might be able to infer plausible rules that the model could be following. For example, Krening et al. (2016) has one reinforcement learning model select action according to a reward structure (which defines its policy) and a second model maps the relations between actions and model states to a lexicon of user-generated terms. This creates an output, from the second model, of best-fit approximations of user terminology to (first) model performance. While this might not give a full and precise account of the activity of the first model, it produces output that is human-understandable. One can also use reinforcement learning to model aspects of human decision-making (*Section 3*).

Figure 18 shows a simulated credit card fraud detection task: analysts are presented with information relating to credit card transactions, e.g. amount, CVC, location, etc. These are either in the form of text displays or red / green alerts. A reinforcement learning model

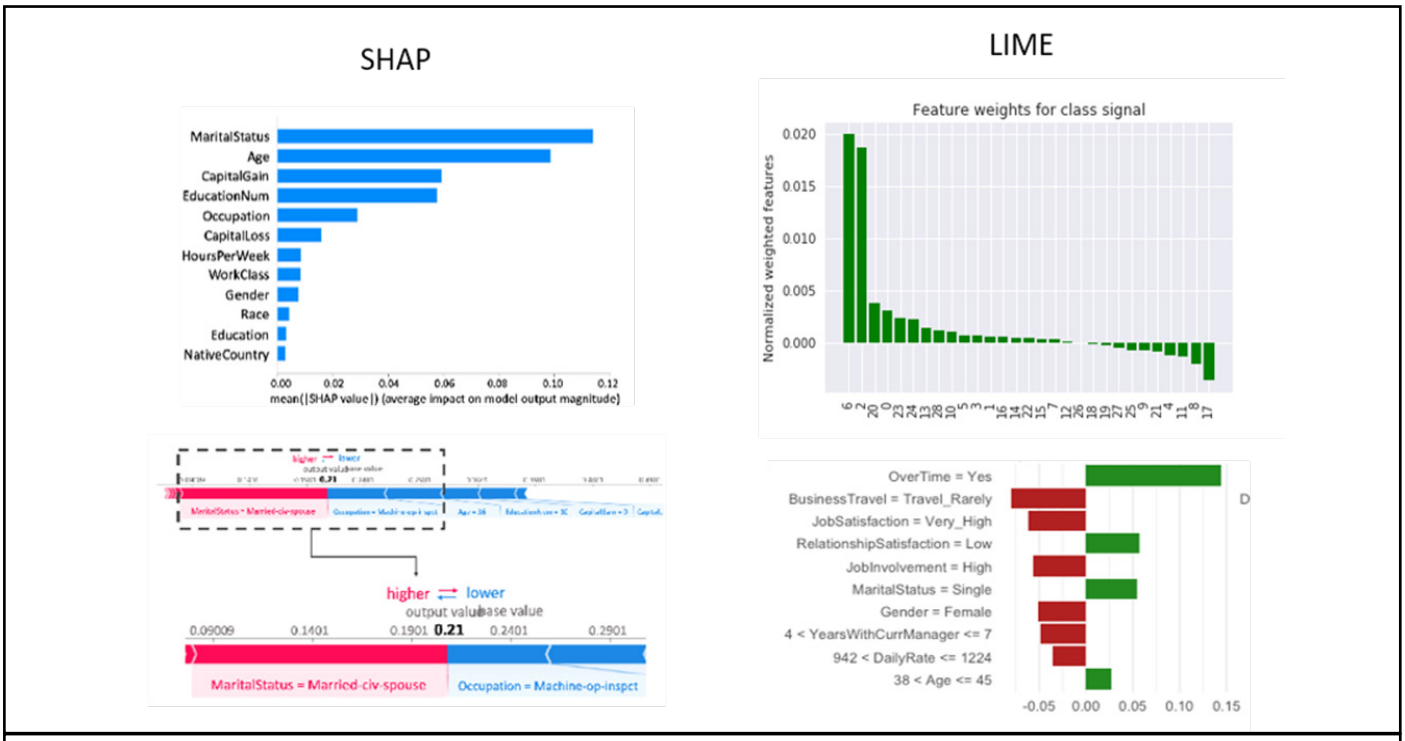


Figure 17. Visualisations output by the implementation of the SHAP Python package (middle), and LIME (right). Top row: global explanations. Bottom row: local explanations [from Kaur et al., 2020]

was trained (based on timings for eye movement and fixations) to attend to regions of the display with a reward given for correct identification of a fraud / no fraud. The task was repeated with human participants and, while the number of cues fixated (left) differ between human data and model, the trends are remarkably consistent. Further investigation suggests that the model responded to similar cues to the human and learned to identify fraud types in a manner comparable to human analysts. This (and the Krening et al., 2016 study) suggest a promising line of research in terms of using AI to model decision strategies so that specific features that are integral to a decision can be identified.

Alternatively, local explanations, i.e. accounts which reflect specific instances, can be created as saliency maps (Wang et al., 2015; Greydanus et al., 2018). *Figure 19* shows examples in which an agent (trained using reinforcement learning) responds to features (in Atari video games). The resulting saliency map shows when and how the agent responded, which can be used to infer the strategy that is being applied. While this need not directly and completely reflect the policy (in terms of the relationship between action and reward that the agent is learning) it can allow the human analyst to form beliefs as to how the agent might behave in similar circumstances.

By way of overview of these different approaches, we can consider them in terms of the situation awareness levels.

- **Level 1** – The human interprets the features that the algorithm is using and forms beliefs to explain the relevance of these features. This involves indicating the data, or features in the data, that the algorithm is using. This can allow the human to infer beliefs around the output (Lipton. 2016; Ribeiro et al., 2016), or the use of saliency maps to allow the human to infer the algorithm's rules (Wang et al., 2015; Greydanus et al., 2018), or the use of belief-desire-intent, especially in robotics (Harbers et al., 2012; Broekens et al., 2010), to explain why behaviour changes as the situation changes (Floyd and Aha, 2016; Lomas et al., 2012), or to present the plan that is being followed (Borgo et al., 2018; Chakraborti et al., 2019; Sreedharan et al., 2018) towards a goal or purpose.
- **Level 2** – The algorithm presents its beliefs, i.e. the rules and principles that it is applying to the feature in a human understandable form. This could involve explanation by simplification, using LIME (Ribeiro et al., 2016), or user-defined abstract features (Kim et al., 2017), or the algorithm offering a policy explanation (Miller et al., 2019), or presentation of model predicates to fill gaps in

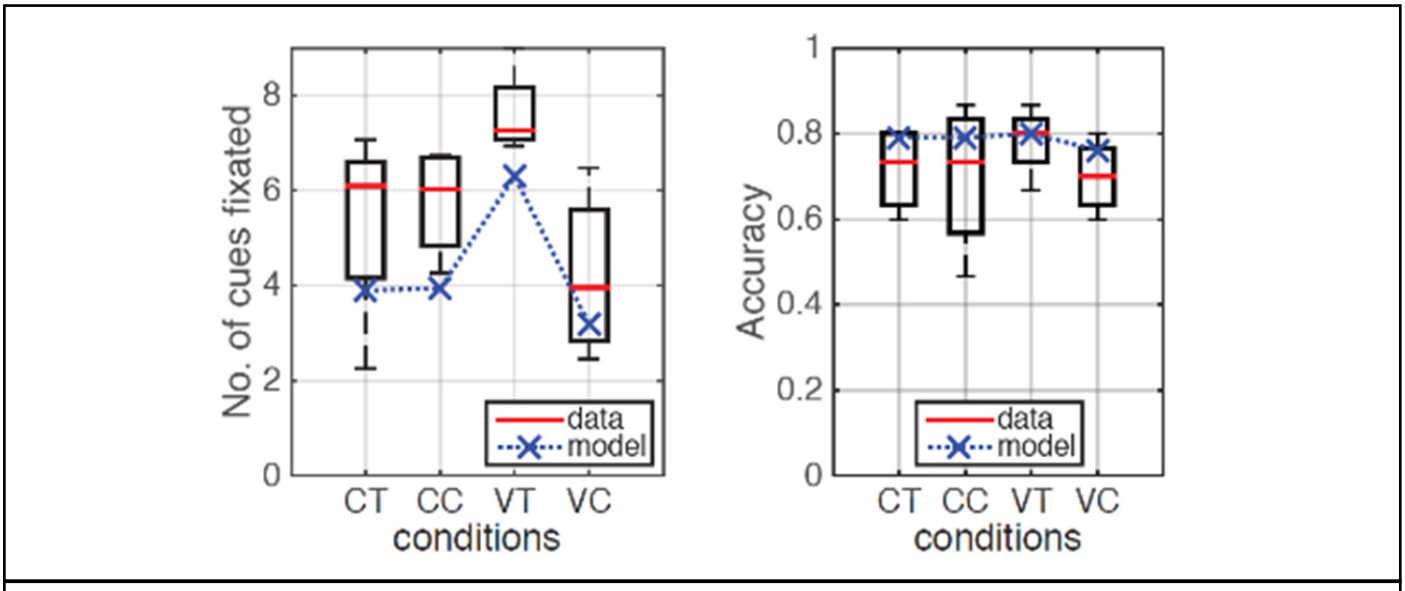
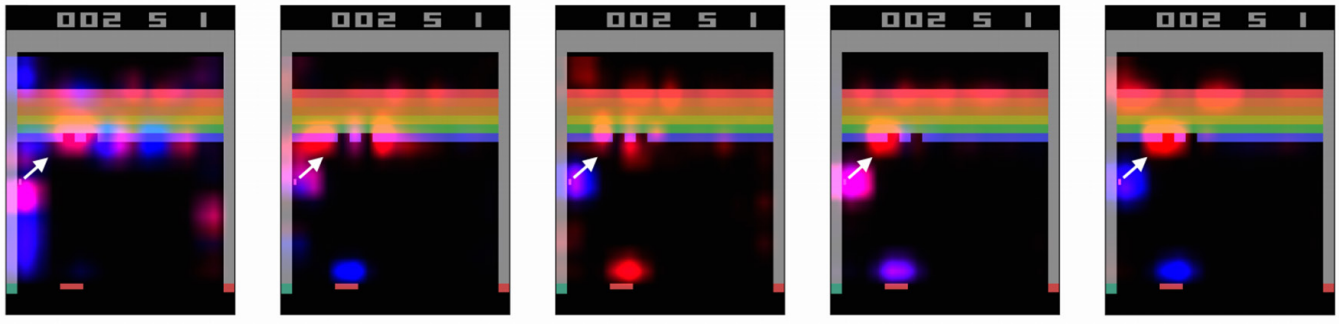
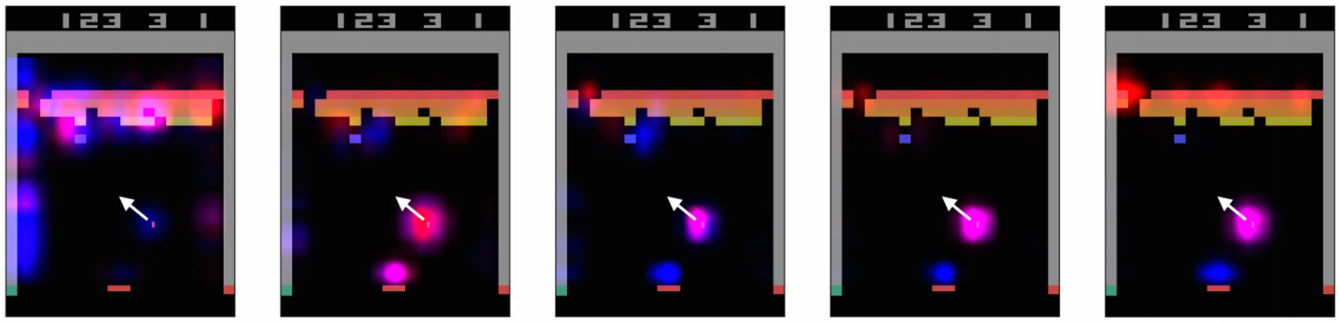


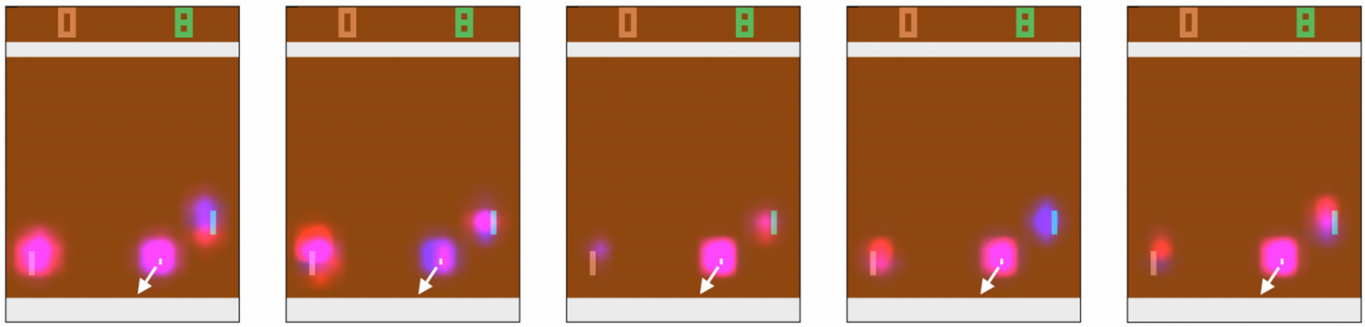
Figure 18. Comparing human and model performance in fraud detection task [Chen et al., 2017]



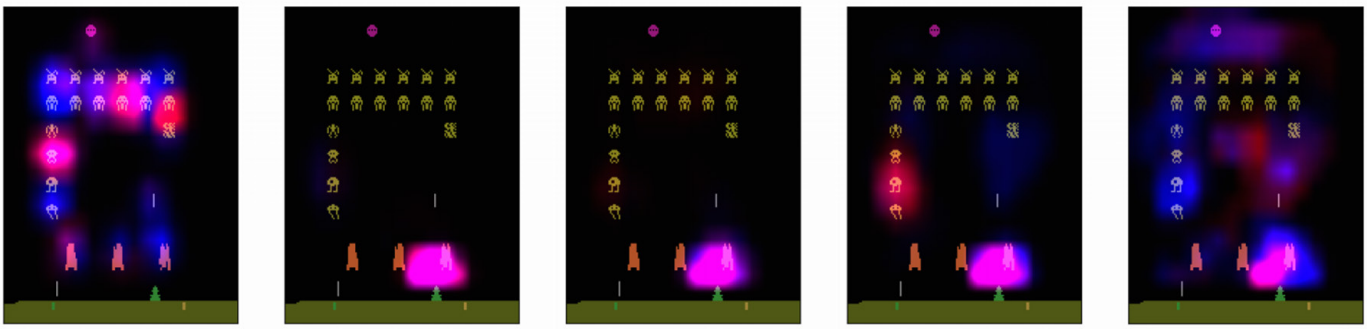
(a) Breakout: learning what features are important



(b) Breakout: learning a tunneling stategy



(c) Pong: learning a kill shot



(d) Space Invaders: learning what features are important and how to aim.

Figure 19. Saliency maps for Atari game play [Greydanus et al., 2018]

user knowledge through use of counterfactual foils (Borgo et al., 2018; Sreedharan et al., 2018).

- **Level 3** – The algorithm can show how it is choosing a decision option from others (Amir and Amir, 2018), or provide a global explanation from combining local instances (Ribeiro et al., 2016), or can modify its decisions in light of human feedback (Holzinger et al., 2018).

THE USEFULNESS OF INTERPRETABILITY TOOLS

While interpretability tools could help people gain insight into the algorithms, these do not always guarantee a useful explanation. An inappropriate explanation could also lead people to over-trust or mistrust model. For example, when the explanation is not faithful and simplifies a model's inherent complexity too much. Although the explanation might be more understandable and convince people more, this could contradict the underlying model (Kaur et al., 2020; Yang et al., 2020; Smith-Renner et al., 2020). If the surrogate model, for instance, is built on a rare or highly specific local instance, then there might be a risk of generalising from this to other instances.

The explanation could also interfere with the analyst's own sensemaking. For example, people might superficially accept an explanation that is presented visually just because it looks more intuitive (Kaur et al., 2020). As a result, a good explanation should balance the fidelity (and validity) of the underlying algorithm and the understandability for people.

Ribeiro et al. (Ribeiro et al., 2016) offered three ways to obtain the utility of explanations. They defined the most significant features in classifiers and calculated their recall rate. Since high feature recall rate alone does not assure that users get insight into the model, they also used subjective rating of trustworthiness of the algorithm. This involved giving users a test set containing outcomes that were known to be wrong, and then defining the rules to produce outcomes that users classify as untrustworthy. Next, they compared these untrustworthy rules with rules which obtained high ratings of trustworthiness by F1 score.

Schmidt and Biessmann (Schmidt and Biessmann, 2019) deigned a quality score for interpretability, based on an information transfer rate (ITR). In this, $I(\hat{Y}_H, \hat{Y}_{ML})$ denotes the mutual information between $\hat{Y}_H$ the annotations provided by human labellers (they were only shown the explanation), and the model predictions $\hat{Y}_{ML}$. $I(\hat{Y}_H, \hat{Y}_{ML})$ could be seen as an objective evaluation of the IT fidelity.

$$ITR = \frac{I(\hat{Y}_H, \hat{Y}_{ML})}{t}$$

In summary, we can evaluate the quality of ML by feature recall rate and people's error-detection capability of the models (in terms of ratings of trust). Kay et al. (2015) sought to improve the F1 score by introducing a new measure, acceptability of accuracy, which is a mixed-effects Bayesian logistic regression against three different weighted power means of precision and recall (harmonic, geometric, and arithmetic). Lim and Dey (2011) have explored similar questions around how much uncertainty is acceptable, how much accuracy is sufficient, and how to best mitigate the uncertainty. Papenmeier et al. (2019) show that accuracy is more important for user trust than explainability, and users cannot be tricked by high-fidelity explanations into trusting a bad classifier. Poursabzi-Sangdeh et al. (2018) varied interpretability by two factors: the number of input features and the model transparency (clear or black box). They showed that increased transparency hampered people's ability to detect when the model makes a sizable mistake and correct for it, seemingly due to information overload. Even more surprisingly, contrary to what one might expect when manipulating interpretability, they found no improvements in the degree to which participants followed the model's predictions when it was beneficial to do so. Their counterintuitive results suggest that users are bad at differentiating the quality of model if model designers only offer unorganised raw features.

One implication of the Poursabzi-Sangdeh et al. (2018) study is that users can struggle with information overload. This is also true when it comes to presenting explanations. For example, Narayanan et al. (2018) varied information load by manipulating the length of explanation, the number of cognitive chunks in explanation, and the number of times the input conditions appeared in explanation. Their result showed that greater complexity results in higher response times and lower satisfaction.

Whether explanations are human-meaningful (or, in other words, whether the feature content is related to the problem) can significantly affect perception of a system's accuracy independent of the actual accuracy observed from system usage (Nourani et al., 2019). Their empirical result implies the importance of the relatedness of feature to the problem by participants significantly underestimating the system's accuracy when it provided weak, less human-meaningful explanations.

Many end-users such as clinicians believe that carefully designed visualisation and presentation can facilitate further understanding of the model (Tonekaboni et al., 2019; Chatzimparmpas et al., 2020; Kaur et al., 2020). However, visualisations can also lead users to believe that a model has higher transparency and intelligibility than it might actually have (Kaur et al., 2020) and that poor visualisations can result in lower levels of trust, even if the model is accurate (Yang et al., 2020).

APPENDIX C: APPLYING THE FRAMEWORK TO A SPECIFIC USE CASE

As noted previously, the workshops produced a series of process flow models that illustrated the ways information is processed in the different scenarios experienced by attendees. In this section, we use a version of financial compliance investigation of possible insider trading on the stock market. Financial trading is heavily regulated and can carry severe penalties, both for those individuals prosecuted and the companies that employ them.

Having produced an initial process model (the red nodes in *Figure 20*), we consider what information might be relevant at each stage (the grey nodes in *Figure 20*). This provides an initial sketch of the process and the information that is used. From this, we construct *Table 9,* where the information is considered in terms of whether it is a feature (i.e. a discrete item or source of information), a cluster (i.e. a combination of features), a belief (i.e. a rule or covering law that defines the cluster of features), or a policy (i.e. a regulation or principle that associates belief with specific action). *Table 9* also introduces the stakeholders who might be involved in this process and indicates whether we assume that these will relate to the different information types.

		Stakeholders			
Information type	**Examples**	**Analysis**	**Internal**	**External**	**External$_b$**
Features	Alerts Individual {position, trades, order, price...} Market {price, movement…} Open Source {news reports, stock market information…}	Compliance Officer Leak / Conspirator Portfolio manager	Compliance Team		
Cluster	Trader Activity	Alert engine			
Beliefs	Statistical anomaly Normal trader activity Market performance Unusual or suspicious activity	Alert criteria	System support Compliance Authority		
Policy	Case Review process Financial Misconduct regulations Compliance Board	Legal team	Senior Compliance Board Company Board Marketing / PR	FSA SEC Shareholders	PR firm Media Investors

Table 9. Financial compliance example

This example provides an initial outline of when and where explanation might be necessary. For example, when information is passed between stakeholders or when specific information types are being considered. For the latter, this might involve justifying the inclusion (or exclusion) of features or the definition of a belief (in terms of the rules / covering law that is being applied in the investigation). For the former, this might involve the translation of information types – either to protect confidentiality or to cater for differences in knowledge between stakeholders. These different forms of explanation are considered by applying our explanation framework to this example.

EXPLANATION TYPES IN THE FINANCIAL COMPLIANCE EXAMPLE

In the financial compliance example (*Table 9*), we will assume that the explainer, X_1, is the compliance officer who is conducting an investigation. The explainee, X_n, could be any of the stakeholders in *Table 9,*.

As noted previously, ***how*** the explanation is couched and presented, by X_1, will depend on which of the stakeholders will be the explainee, e.g. in terms of how much of the set of features are to be shared (as some of this might either be commercially sensitive or a matter of conjecture as the investigation unfolds), or in terms of the how much of the belief is shared (as some of this might be specific to the company or might highlight particular investigatory practices which are sensitive).

S1 ≈ S2 AND R1 R2

In the financial compliance example (*Table 9*), the first instance, X_1 and X_2, might be two compliance officers (who have access to the same set of features and share the same beliefs (knowledge) of the investigation process) and the same understanding of policy.

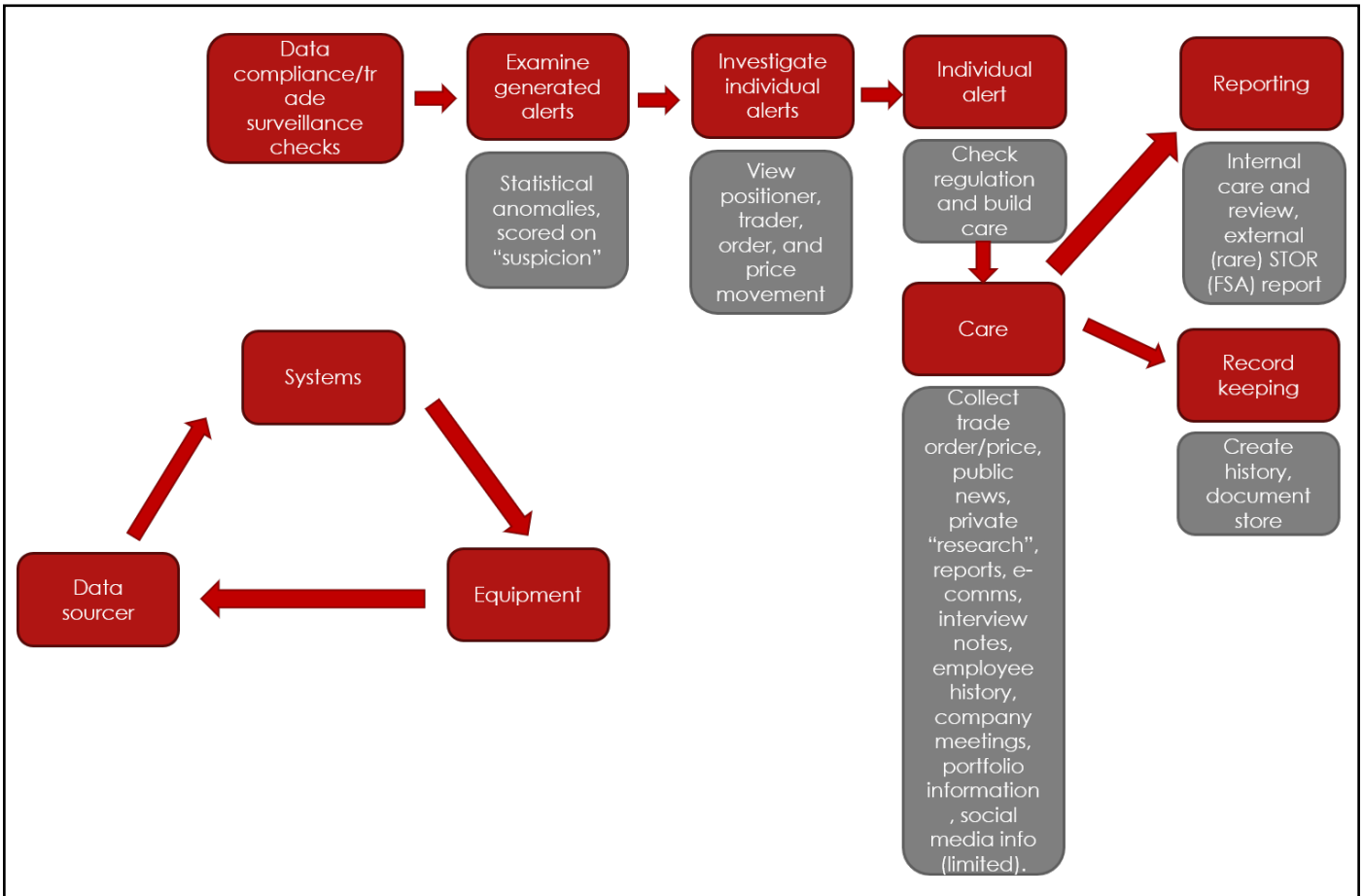


Figure 20: Saliency maps for Atari game play [Greydanus et al., 2018]

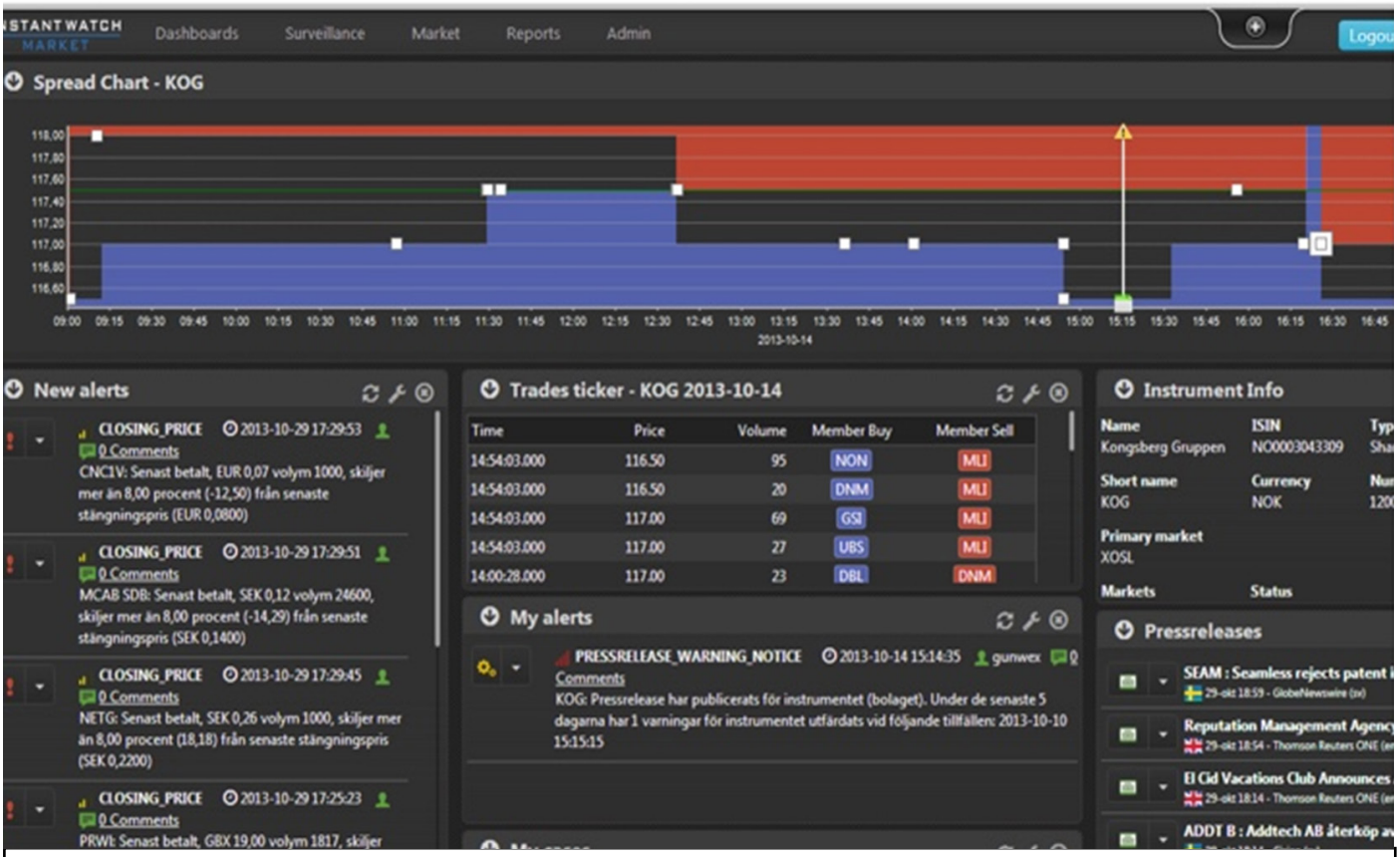


Figure 21. Monitoring trader activity [https://www.trapets.com/services/instantwatch-market/]

In this case, situation awareness and relevance can be assumed to be common, and any explanation involves the exchange of information to ensure that both are on the same page (i.e. monitoring that their respective knowledge bases sufficiently overlap).

SX1 ≈ SX2 AND RX1 RX2

If there is disagreement between X1 and X2 (either, for example, because the compliance officers draw different conclusions or because the compliance officer is outlining the case to the trader as part of an initial investigation), the features will be shared (e.g. in the form of a set of evidence) and interpretation of these (in terms of whether this constitutes lack of compliance) involves a mismatch in terms of relevance. In this instance, the purpose of the explanation is to seek alignment between X1 and X2 in terms of relevance, i.e. whether there is a compliance case to answer.

SX1 ≈ SX2 AND RX1 RX2 AND ΔRX2 ≈ RX1 RX1

Alignment, as mentioned previously, does not mean there should be absolute agreement in R. In this instance, the aim would be to ensure that, perhaps through further sharing of situation features, X1 and X2 agree on the feature set to be used for the analysis and on the grounds for the investigation. In this case, R describes the basis of the case (even if there is not agreement that the evidence supports this).

S1 ≠ S2 AND R1 R2 AND ΔR2 ≈ R1 R1 AND A2 = S2

Consequently, the initial investigation might be to encourage a change in the behaviour of X2. For example, in a compliance investigation, an issue might arise from a lack of record keeping and so an outcome might be for X2 to appreciate how specific features of the situation can be interpreted in terms of specific beliefs, such that paying more attention to the

interpretation and recording of these features could reduce the possibility of investigation in the future.

SX1 ≠ SX2 AND RX1 RX2

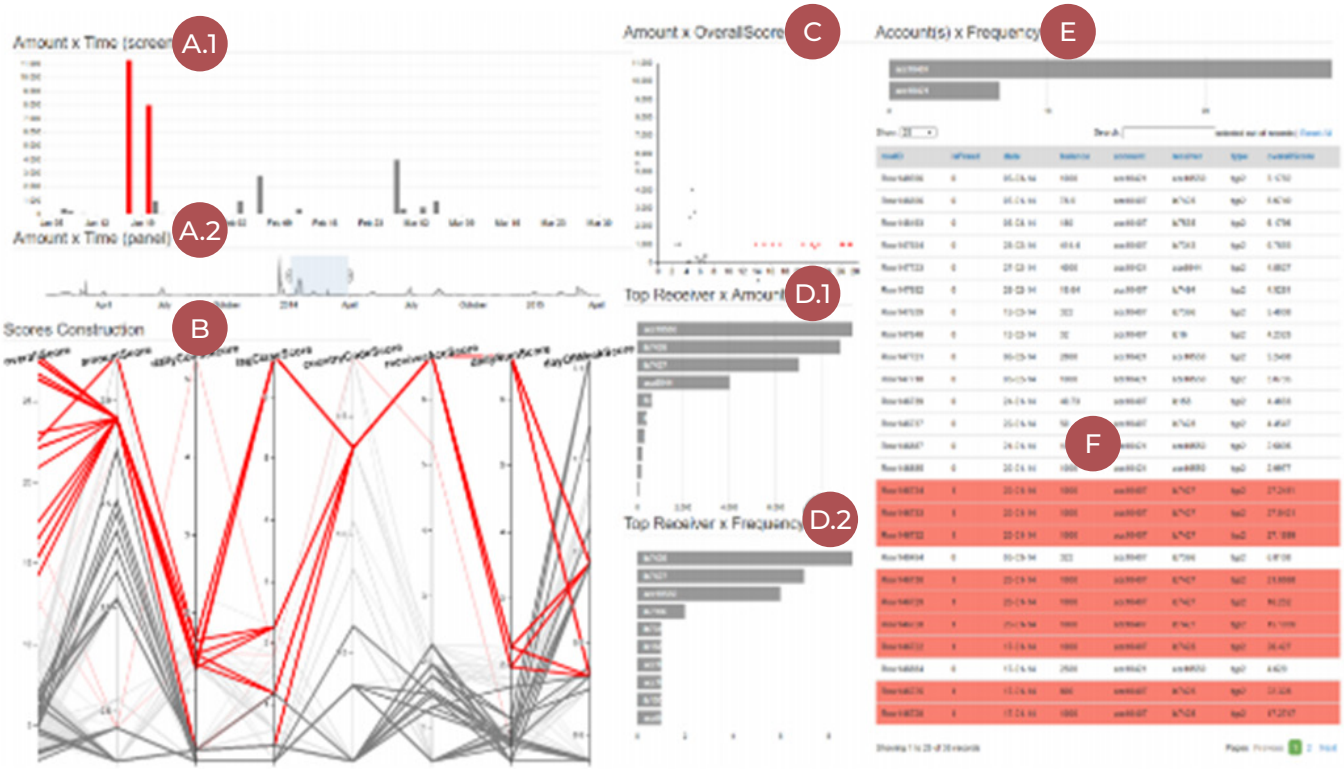
Where there continues to be misalignment between X1 and X2, the explanation would build a detailed case to demonstrate how specific sets of features can be interpreted in terms of specific beliefs and / or policy. This type of scenario is more likely to occur between an X1 and X2 who come from different backgrounds and, therefore, have different knowledge sets i.e. a compliance officer and the company lawyer for instance.

SX1 ≠ SX2 AND RX1 RX2

From the further investigation, while both parties might share the beliefs and understanding of policy, they might disagree on the definition or interpretation of features.

VISUALISING MACHINE LEARNING EXPLANATIONS

We briefly consider the role that ML might play in this example through the different information types (as these apply to examples of financial decision-making). In this section, we consider a variety of financial decision-making (from compliance analysis to loan underwriting). The intention is to provide examples of the types of user interface (rather than algorithms) that



Screenshot of EVA (Event detection with Visual Analytics). (A.1, A.2) Temporal Views: a filler was applied in (A.2) to the period from January 2014 until April 2014). (B) Score Construction View: each line represents a transaction and its scores. (C) Amount vs Overall Score Scatterplot. (D.1, D.2) Ranks of accounts that received the highest amounts of money from the selected account. (E) Accounts Selector: bars shows the amount of transactions from each account. (F) Dynamic table of raw transaction data. In all views, elements that represents suspicious data are highlighted in red.

Figure 22. Event detection with visual analytics [Leite et al., 2017]

Decision Explanation Illustrator

	Application 1	Application 2	Application 3	Application 4	Application 5	Application 6
Factual Rule Base						
Affordability test	FAIL	PASS	PASS	PASS	PASS	PASS
Number and amount of bankruptcy	NONE	>=1	NONE	NONE	NONE	NONE
Number of IVA & CCJ	NONE	NONE	NONE	NONE	NONE	NONE
Number of payday loans	NONE	>=1	NONE	NONE	NONE	NONE
Decision: Automated Application Acceptance or Decline	DECLINE	DECLINE	ACCEPT	ACCEPT	ACCEPT	ACCEPT

Heuristic Rule Base	F	R	F	R	F	R	F	R	F	R	F	R
Unsecured loans	*Application declined in early stage*		*Application declined in early stage*		0.79	0.21	0.64	0.36	0.27	0.73	0.64	0.36
Secured loans					0.64	0.36	0.74	0.26	0.74	0.26	0.75	0.25
CCJ, IVA, Bankruptcy & payday loans					0.76	0.24	0.66	0.34	0.76	0.24	0.66	0.34
Searches					0.78	0.22	0.76	0.24	0.81	0.19	0.62	0.38
Credit score					0.79	0.21	0.61	0.39	0.62	0.38	0.73	0.27
Loan criteria, property valuation & property type					0.95	0.045	0.04	0.96	0.04	0.96	0.88	0.12
Predicted Output					**0.96**	**0.04**	**0.07**	**0.93**	**0.07**	**0.93**	**0.89**	**0.11**
Decision: Fund or Reject	REJECT		REJECT		FUND		REJECT		REJECT		FUND	
Textual explanation for a rejected application	Application has failed affordability test		Applicant have inadequate number and amount of bankruptcy and payday loans				-The property is poor and it has failed mortgage valuation. -The loan application do not fit our product-plan (loan criteria).		-The applicants have bad unsecured loan. -The property is poor and it has failed mortgage valuation. -The loan application do not fit our product-plan (loan criteria)			

Figure 23. Illustrating beliefs in loan underwriting [Sashan et al., 2020]

have been proposed to support decision-makers. By analogy, we suggest that variants of the different user interfaces could be considered to support the role of the compliance officer in the use case.

FEATURES

By displaying key features that are relevant to trading alerts, *Figure 21* provides a dashboard that an analyst can use to interpret alerts (listed on the bottom-left of the screen).

The implication of displaying features is that this will allow the compliance officer to quickly ascertain the key features which led to an alert being raised. In this case, the explanation is seeking to ensure that the situation seen by the automation that generated the alert aligns with that seen by the compliance officer. This would mean that S1 ≈ S2. If one assumes that the definition of the feature sets (***S***) follows similar rules, then this could also imply that R1 ≈ R2. However, the

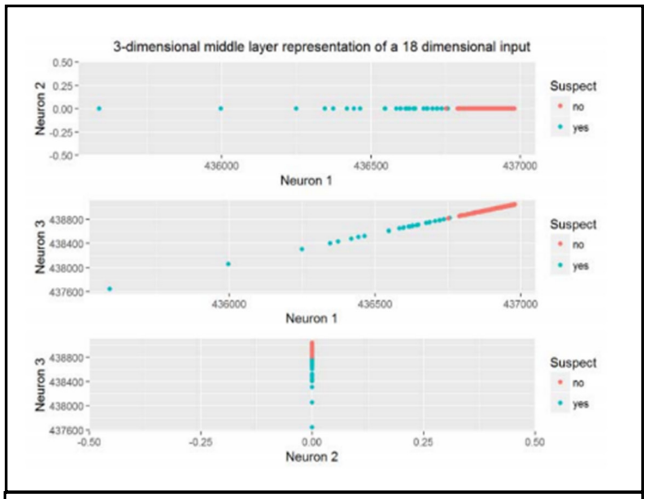


Figure 24. Discriminating fraudulent from nonfraudulent transactions [Paula et al., 2016]

display solely of features does not allow the analyst to either interrogate the underlying beliefs that led to the automation raising an alert or to question this belief. Moreover, the display is the motivator for further investigation (probably drawing on other information sources) that will result in the manual production of a report that summarises the investigation.

CLUSTERS

Providing a dashboard with key features requires the analyst to consult additional sources to build up a composite profile of the activity. Elaborating the dashboard to create clusters of these features (*Figure 22*) could be useful in identifying patterns and trends.

Collating features into charts and tables provides the analyst with a summary that can be interpreted in terms of rules. Indeed, an experienced analyst would most likely recognise recurring patterns across different instances. That is, similar activities might produce clusters that have similar visual appearance, such that there would be fingerprints that the analyst picks up on that corresponds to specific activities. In this way the alert relates not only to specific features but to the groupings of these features. While this might aid recognition-primed decision-making (Klein et al., 1989), it does not provide access to the underling rules used to generate the clusters.

BELIEFS

In *Figure 23*, the rules that are used to reach a decision are listed, together with an indication of whether the rules have been met or breached (with pass / fail, colour coding, accept / decline in the top of the display). Additional information, in the form of weighting for contributing features, is presented in the next part of the display, together with recommendation for fund / reject. In this way, the automation's rules are exposed to the human decision-maker. The textual explanation, at the bottom of the screen, is clear and concise. In this instance, we can claim that the user interface is intended to support $Sx1 \approx Sx2$ and $Rx1 \neq Rx2$ and $Rx2 \approx rx1 \subseteq Rx1$ (where the automation, X1, is providing sufficient information to allow the analyst, X2, to agree with the definition of relevance that is being applied in each application). What is not apparent here is whether the human is able to apply counterfactuals to the decision. For example, if we consider Application 4, the heuristic rule base identifies 'loan criteria, etc.' as below criteria, but what might happen if the applicant was able to amend this?

POLICY

Paula et al. (2016) describe an autoencoder with three hidden layers (the middle one, shown in *Figure 24*, has three neurons). Using a training set of 20 fraudulent transactions defined as 'yes' and 'no' shows that the middle layer was capable of linear separation of these. While this does not tell us *how* the autoencoder is making its decisions, this does provide an indication of how the policy that it is discovering is being applied. In this instance, the aim is to support, to some extent, $Sx1 \neq Sx2$ and $Rx1 \approx Rx2$ where it is recognised that the automation, X1, is not focusing on the same features as the analyst, X2, but there is an attempt to allow alignment in the definition of relevance.

For more information on CREST
and other CREST resources, visit
www.crestresearch.ac.uk

CENTRE FOR RESEARCH AND
EVIDENCE ON SECURITY THREATS

21-024-01