A Human-Centred Process Model for Explainable AI

Chris Baber, Emily McCormick and Ian Apperly

January 13, 2022

A Human-Centered Process Model for Explainable AI

Chris BABER, Emily McCORMICK, Ian APPERLY
University of Birmingham

ABSTRACT

With the increasing use of inscrutable Artificial Intelligence systems to support human decision makers, there has been much interest in what it means for these systems to provide 'explanation'. In this paper, the concern is with applying a simple formalism that can express a minimal set of features that can be used to define an explanation. It is argued that few contemporary AI systems support this minimal set. Advice is provided on how future developments in explainable AI systems could adhere to this minimal set.

KEYWORDS

XAI; Explainable AI; Explanation.

INTRODUCTION

Developments in Artificial Intelligence (AI) have demonstrated impressive performance, particularly on well-defined domains such as image processing or game playing. However, the techniques that are deployed can be opaque for the human user which raises the question of how AI systems can provide explanation (Neerincx et al., 2018; Rosenfeld and Richardson, 2019) and there is growing requirement for explainable AI (XAI) in Regulatory frameworks. Having said this, in 2017 Google's research chief, Peter Norvig, pointed out of the irony of expecting computers to provide 'explanations' when humans can be poor at doing this.

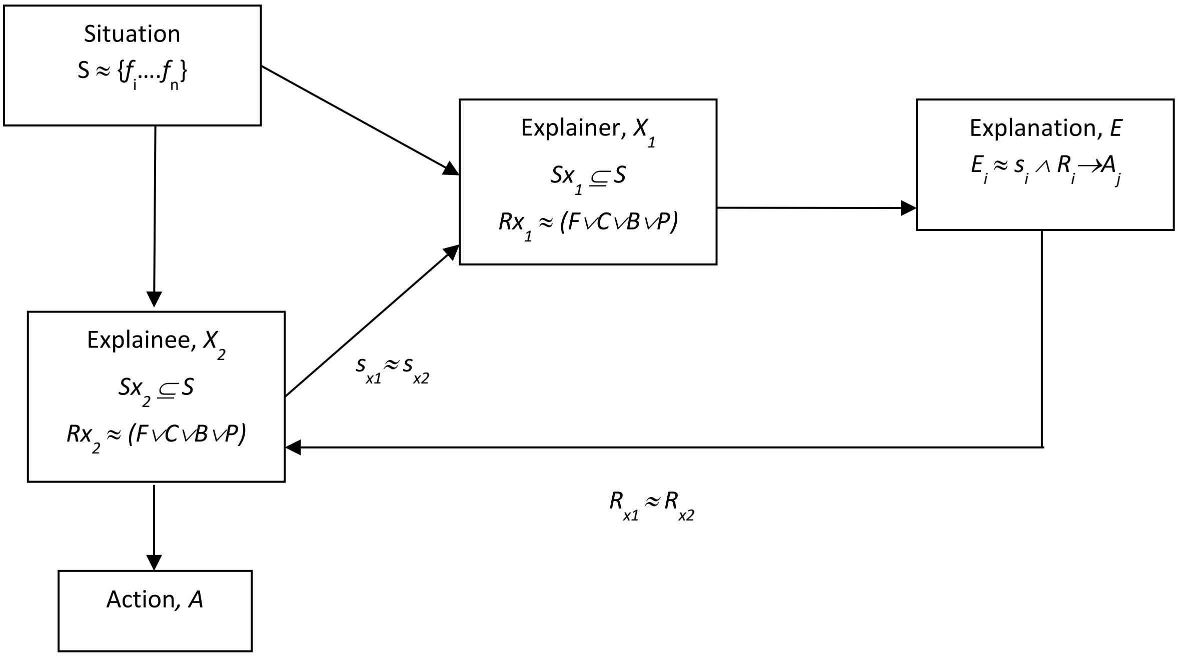
Much of the work on explainable AI (XAI) leans heavily on a computer-centric perspective (Springer, 2019). For example, Holzinger et al. (2020) assume that human and AI system have equal access to a 'ground truth'. From this, explainability *"…highlights decision relevant parts of machine representations…, i.e., parts which contributed to model accuracy in training or to a specific prediction."* In common with much of the XAI literature, this does not provide a role for the human, other than as passive recipient. The implication is that the AI system is able to introspect on its own processes to generate an explanation. The resulting explanation is then presented to the user, with description of the AI system's processes or the features ('decision relevant parts') that it has used. In this way, an explanation is simply a recommendation (from the AI system) plus the features that relate to this. As Miller (2017) notes, a problem with such an attitude is that it is based on the designer's intuition of what makes a 'good' explanation rather than on a sound understanding of how humans respond to, and make use of, explanation. This does not indicate why *some* features were selected or why the recommendation is appropriate to the user's concerns. Nor does it situate explanation in the wider organisation; it is likely that an explanation for the analyst will be distinct from that for the person managing data collection or the manager who will be briefed by the analyst.

For Holzinger et al. (2020) aspects of the situation (defined as a ground truth) are combined into a statement; that is, the explanation is simply an expression of this statement. This implies that there is a linear interpolation from features to explanation. This is similar to Hempel and Oppenheim's (1948) 'Covering Law Model' which was concerned with the ways in which Historians might explain an Event in terms of antecedent Causes. However, 'ground truth' (assumed by Holzinger's process model and by the covering law model) is seldom fully defined (leading to ambiguity in the selection of relevant features). This means that simply stating the situation aspects without an indication of why these (rather than other aspects) were selected might not lead to a useful or usable explanation.

Hoffman et al. (2018) provide a comprehensive review of literature relating to explanation. From this review, explanation involves sensemaking by the human (to contextualise the output of the AI system) and we agree that an appropriate framework for considering this is the Data-Frame model of sensemaking (Klein et al., 2007). Further, sensemaking (and its relationship with explanation) relies on the recognition that the process (of providing and receiving an explanation) must be reciprocal, iterative, and negotiated. This process relies on 'explainer' and 'explainee' reaching alignment. In other words, explanation involves 'common ground' (Clark, 1991) in which there is sufficient alignment in understanding for conversation to proceed. The nature of the conversation will depend on the situation in which the explanation is being provided and the goals of the explainee. For example, the explainee might be a 'trainee' who seeks to understand the explanation to learn criteria for a decision or might be an 'analyst' using the recommendation from the AI system to apply as a policy.

A PROCESS MODEL FOR EXPLANATION

Figure 1 illustrates the relationship between an Explainer and an Explainee in a Situation (Baber et al., 2020). A Situation has *features* which are analogous to the notion of 'data' in the Data-Frame Model. We use the term 'features' (rather than data) because the word 'data' has a narrow definition in the AI literature. Notice that in figure 1, relations are indicated by ≈ to indicate that these relations are partial, provisional and approximate and, just as the Data-Frame Model emphasises, require continual monitoring, checking and refinement. While the Data-Frame Model uses the term frame, this also has a privileged meaning in the AI literature. So, we adopt the term Relevance (Sperber and Wilson, 1982) to refer to the rationale for why features are selected by explainer or explainee. So, relevance could be defined by one or two features, F, or a cluster of features, C, or a belief, B (which allows predictions to be made about Features, Clusters and Situations), or a Policy, P (which associates Actions with the Situation). It is important to note that 'relevance' is relative, i.e., the definition of relevance would depend not simply on the features in the situation but on the prior experience of the people and their goals; the same situation could result in different Situation Models for the people experiencing it.



Situation
$S \approx \{f_i \ldots f_n\}$
Explainer, X_1
$Sx_1 \subseteq S$
$Rx_1 \approx (F \vee C \vee B \vee P)$
Explanation, E
$E_i \approx s_i \wedge R_i \rightarrow A_j$
Explainee, X_2
$Sx_2 \subseteq S$
$Rx_2 \approx (F \vee C \vee B \vee P)$
$s_{x1} \approx s_{x2}$
$R_{x1} \approx R_{x2}$
Action, A

Figure 1. Defining Explanation

The Explainer and Explainee attend to features in a Situation (i.e., their 'situation model', *S*, consists of features which are selected as relevant to each person). From this, an explanation *could* involve overlap between 'situation models'. That is, *sx1≈sx2*. For models like that of Holzinger et al. (2018) the process of 'alignment' seems to be one-way (from computer to person) and assumes that this will result in the person adopting the same situation model as the computer. But, for a human-centered approach, this does not feel plausible. Rather, there needs to be scope for the alignment of situation models to be open to negotiation and dispute. We assume that explanations will involve second-order situation models, i.e., accounts which, at a later time, summarise the situation to another explainee. This means that, in order for 'situation models' to make sense to both parties, there needs to be alignment between definitions of 'relevance'. In this way, the primary means by which explanation operates is through alignment of 'situation model' *and* 'relevance' held by explainer and explainee, often (but not always) to encourage an Action on the part of explainee. However, we agree with Lipton (2016) that one must not treat 'explanation' as a monolithic concept. That is, different Situations (and different explainers and explainees) will require different types of explanation.

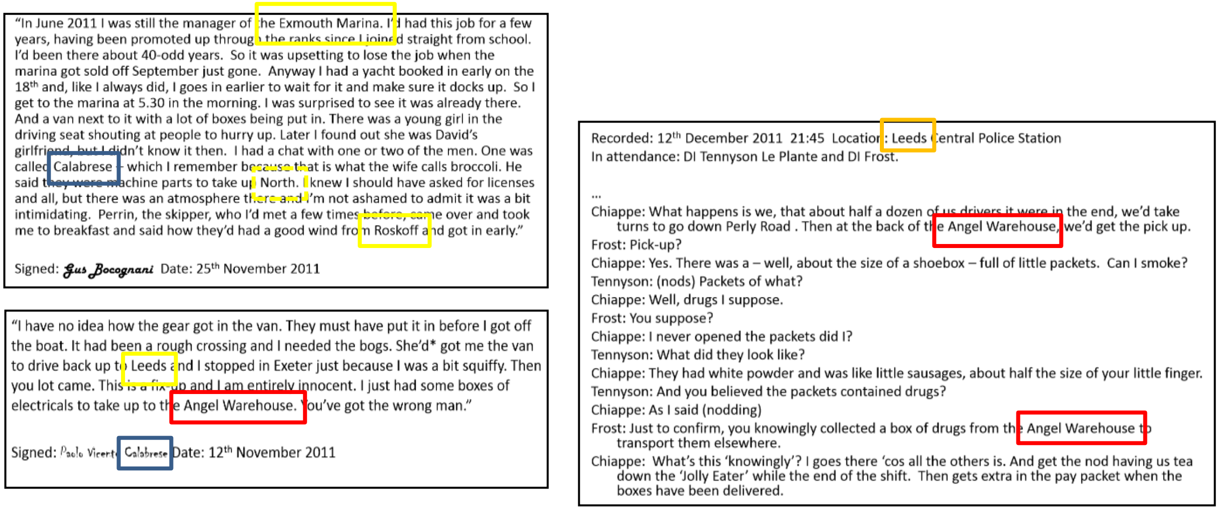
TYPES OF EXPLANATORY DISCOURSE

Implicit in the dichotomy, of situation model plus relevance, is the assumption that either of these can be 'surfaced' (i.e., brought to conscious awareness and expressed in words). Surfacing features of a Situation Model that are Relevant means not only an ability to introspect on our cognitive processes but also an ability to put the tacit knowledge that this implies into words. Further, AI system (particularly 'deep' AI) will be unable to introspect on its own processes. But explanations between humans seems to recognise this problem and we have techniques for managing explanatory discourse that enable us to challenge and explore this (Miller, 2017) – and these techniques have not been commonly applied to XAI (Miller et al., 2017).

We assume four types of explanatory discourse in which Situation Model or Relevance are Aligned or Challenged:

	Align	Challenge
Situation Model	Explainer draws attention to specific features in the situation.	Explainee disputes the indicated features and requires clarification of the situation model being applied.
Relevance	Explainer presents underlying rationale for the situation model.	Explainee offers alternative definitions of relevance or appeals to 'counter-factual' (or 'what if') examples, e.g., what if a given feature was present or absent.

To illustrate these types of explanatory discourse in human activity, figure 2 shows evidence used in the *North x South-West Exercise* for Intelligence Analysts (Baber et al., 2015, 2016). Highlighted features (within boxes) require knowledge of UK and Bretagne geography. Thus, to 'explain' the link between the pieces of evidence in figure 2, explainer and explainee need to agree on geography, e.g., the 'Angel Warehouse' is in Leeds, Leeds is 'up North', Exmouth and Leeds are connected by road, Roskoff and Exmouth are separated by an expanse of water.



"In June 2011 I was still the manager of the Exmouth Marina. I'd had this job for a few years, having been promoted up through the ranks since I joined straight from school. I'd been there about 40-odd years. So it was upsetting to lose the job when the marina got sold off September just gone. Anyway I had a yacht booked in early on the 18th and, like I always did, I goes in earlier to wait for it and make sure it docks up. So I get to the marina at 5.30 in the morning. I was surprised to see it was already there. And a van next to it with a lot of boxes being put in. There was a young girl in the driving seat shouting at people to hurry up. Later I found out she was David's girlfriend, but I didn't know it then. I had a chat with one or two of the men. One was called Calabrese – which I remember because that is what the wife calls broccoli. He said they were machine parts to take up North. I knew I should have asked for licenses and all, but there was an atmosphere there and I'm not ashamed to admit it was a bit intimidating. Perrin, the skipper, who I'd met a few times before, came over and took me to breakfast and said how they'd had a good wind from Roskoff and got in early."

Signed: Gus Bocognani Date: 25th November 2011

"I have no idea how the gear got in the van. They must have put it in before I got off the boat. It had been a rough crossing and I needed the bogs. She'd* got me the van to drive back up to Leeds and I stopped in Exeter just because I was a bit squiffy. Then you lot came. This is a fix-up and I am entirely innocent. I just had some boxes of electricals to take up to the Angel Warehouse. You've got the wrong man."

Signed: Paolo Vicente Calabrese Date: 12th November 2011

Recorded: 12th December 2011 21:45 Location: Leeds Central Police Station
In attendance: DI Tennyson Le Plante and DI Frost.

…
Chiappe: What happens is we, that about half a dozen of us drivers it were in the end, we'd take turns to go down Perly Road . Then at the back of the Angel Warehouse, we'd get the pick up.
Frost: Pick-up?
Chiappe: Yes. There was a – well, about the size of a shoebox – full of little packets. Can I smoke?
Tennyson: (nods) Packets of what?
Chiappe: Well, drugs I suppose.
Frost: You suppose?
Chiappe: I never opened the packets did I?
Tennyson: What did they look like?
Chiappe: They had white powder and was like little sausages, about half the size of your little finger.
Tennyson: And you believed the packets contained drugs?
Chiappe: As I said (nodding)
Frost: Just to confirm, you knowingly collected a box of drugs from the Angel Warehouse to transport them elsewhere.
Chiappe: What's this 'knowingly'? I goes there 'cos all the others is. And get the nod having us tea down the 'Jolly Eater' while the end of the shift. Then gets extra in the pay packet when the boxes have been delivered.

Figure 2. Items of Evidence for Investigation

As the first step in creating an explanation, Explainer and Explainee should attend to the same features , i.e., $S_{x1} \approx S_{x2}$. So, both people attend to the highlighted sections of figure 2 (if they do not, then the Explainer could point to each of these). But this is not sufficient to guarantee an explanation because the definition of relevance might differ between Explainer and Explainee, i.e., $R_{x1} \neq R_x$. From this, the Explainer wants to change the Explainee's notion of relevance so that it overlaps with part or all of the Explainer's notion of relevance, i.e., $\Delta R_{x2} \approx r_{x1} \subseteq R_{x1}$. Thus, the contents of the 'boxes' that are being transported by yacht and van, e.g., 'machine parts', 'electricals', or 'shoeboxes…of white powder', could be inferred by the Explainer to be the same thing; reference to 'electricals' or 'machine parts' could be deliberately misleading (based on the belief that the real content of the boxes is 'white powder'). The Explainer might point out ambiguity in the definition of 'contents' – in the expectation that the Explainee would recognise this. Or the Explainer might adjust the Explainee's relevance in order to have the Explainee perform an action, i.e., $\Delta R_2 \approx r_1 \subseteq R_1$ *and* $A_2 = \Delta s_2$. Believing that the 'boxes' contain 'white powder' (rather than electrical goods), the Explainer might seek to persuade the Explainee to conduct further analysis, e.g., collect Forensic reports from the boxes that have been recovered, or seek other instances where 'boxes' have been mentioned in interviews or reports, or speak to other people etc.

There will be situations in which the explainer and explainee are not able to reach alignment on the definition of relevance. For example, presentation in Court requires the Explainer to reconstruct the Situation in sufficient detail for the Explainee (in this case judge, jury or barristers) to appreciate (a) the selection of Features and (b) the Relevance of these features to the Situation. Dispute or disagreement could arise if the Explainee does not accept the features or their relevance, e.g., the explanation (of the content of the boxes) rests on the belief that they do not contain 'electricals' or 'machine parts' and that they do contain 'white powder'.

HOW DO AI SYSTEMS SUPPORT EXPLANATION?

Langley (2019) defines the operation of an agent capable of producing an explanation as:

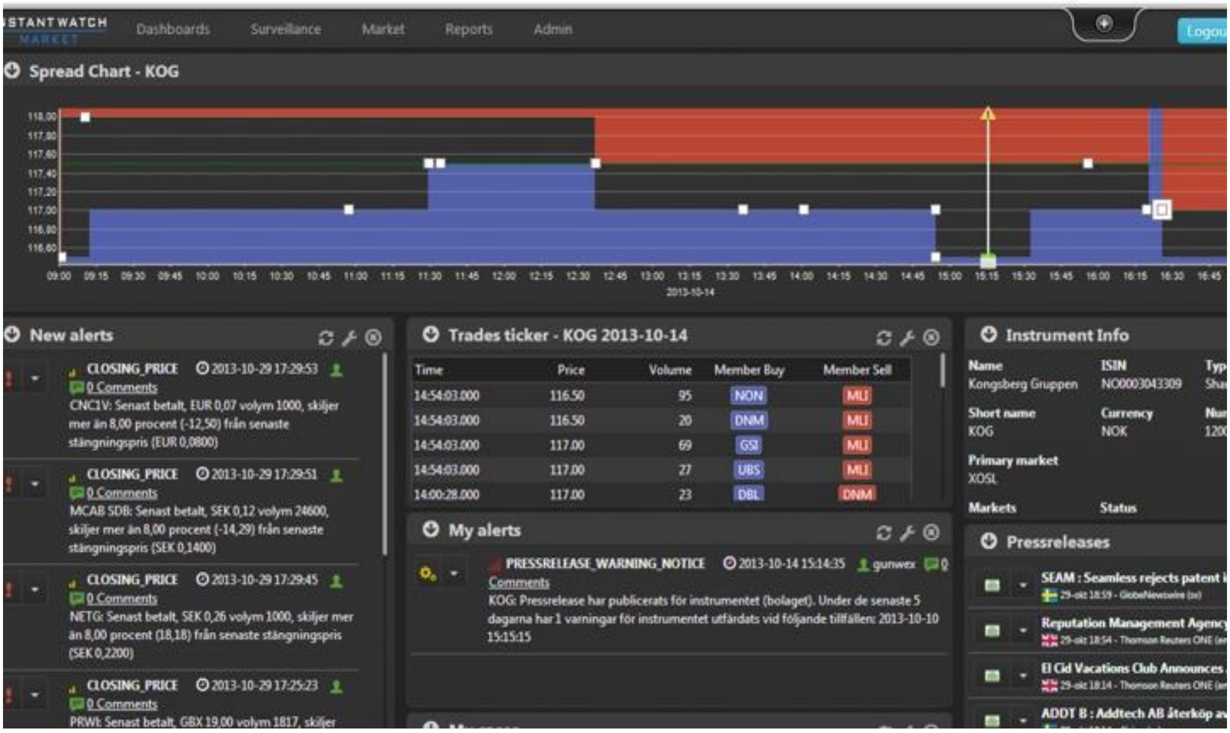
> *"Given: Knowledge defining a space of possible solutions;*
> *Given: Criteria for evaluating candidate solutions;*
> *Given: An annotated search tree that includes solutions found for some reasoning task…;*
> *Given: A query about why a solution ranks above others;*
> *Produce: An explanation why the solution is preferable".*

In terms of our process model, the first two items in this list relate to the Situation Model that is used by the agent, and the second two relate to its definition of Relevance. From Langley's (2019) perspective, the purpose of 'explanation' will be to present the agent's situation model and relevance. But this seems to assume that alignment really means acceptance by the user. So, from this definition, explanation cannot be challenged. In other words, this definition rests on the assumption of transmission of the Explanation to the user rather than an explanatory discourse. Further, while the 'situation models' that humans create might be causal (e.g., in terms of plausible 'causes' of a given event or feature), it is more likely that the models that machines create are relational (e.g., correlation, regression, distance, similarity). This leads to the subtle problem of mistaking correlation for causation, i.e., the human could misinterpret correlations, on which the AI systems depend, for either causal (that is generalisable) relations or predictive beliefs. But neither of these (causal relations or predictions) are integral to the AI's algorithms. Next, we consider examples of how AI systems present explanations.

Features and Clusters

Many approaches to XAI require the user to infer the relevance of specific features to a recommendation. A popular approach to XAI involves explanation-by-simplification. For example, Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) uses a specific instance which concentrates on local fidelity, i.e., the relations of that specific instance. In effect, the approach echoes the logic of the covering law.

Figure 3 shows alerts in financial trading. The implication is that this will allow the trader to ascertain the key features which led to an alert being raised. In this case, the explanation is the AI system's situation model. However, the display solely of features does not allow the trader to interrogate the underlying beliefs that led to the AI system raising an alert or to question its situation model. Moreover, the display is intended to motivate the user to conduct further investigation (probably drawing on other information sources) and, as such, cannot be, of itself, an explanation.



CLOSING_PRICE 2013-10-29 17:29:53
0 Comments
CNC1V: Senast betalt, EUR 0,07 volym 1000, skiljer mer än 8,00 procent (-12,50) från senaste stängningspris (EUR 0,0800)

CLOSING_PRICE 2013-10-29 17:29:51
0 Comments
MCAB SDB: Senast betalt, SEK 0,12 volym 24600, skiljer mer än 8,00 procent (-14,29) från senaste stängningspris (SEK 0,1400)

CLOSING_PRICE 2013-10-29 17:29:45
0 Comments
NETG: Senast betalt, SEK 0,26 volym 1000, skiljer mer än 8,00 procent (18,18) från senaste stängningspris (SEK 0,2200)

CLOSING_PRICE 2013-10-29 17:25:23
0 Comments
PRWI: Senast betalt, GBX 19,00 volym 1817, skiljer

Time	Price	Volume	Member Buy	Member Sell
14:54:03.000	116.50	95	NON	MLI
14:54:03.000	116.50	20	DNM	MLI
14:54:03.000	117.00	69	GSI	MLI
14:54:03.000	117.00	27	UBS	MLI
14:00:28.000	117.00	23	DBL	DNM

PRESSRELEASE_WARNING_NOTICE 2013-10-14 15:14:35 gunwex 0
Comments
KOG: Pressrelease har publicerats för instrumentet (bolaget). Under de senaste 5 dagarna har 1 varningar för instrumentet utfärdats vid följande tillfällen: 2013-10-10 15:15:15

Name: Kongsberg Gruppen
ISIN: NO0003043309
Typ: Sha
Short name: KOG
Currency: NOK
Nur: 120
Primary market: XOSL
Markets
Status

SEAM : Seamless rejects patent i
29-okt 18:59 - GlobeNewswire (sv)

Reputation Management Agency
29-okt 18:54 - Thomson Reuters ONE (en

El Cid Vacations Club Announces
29-okt 18:14 - Thomson Reuters ONE (en

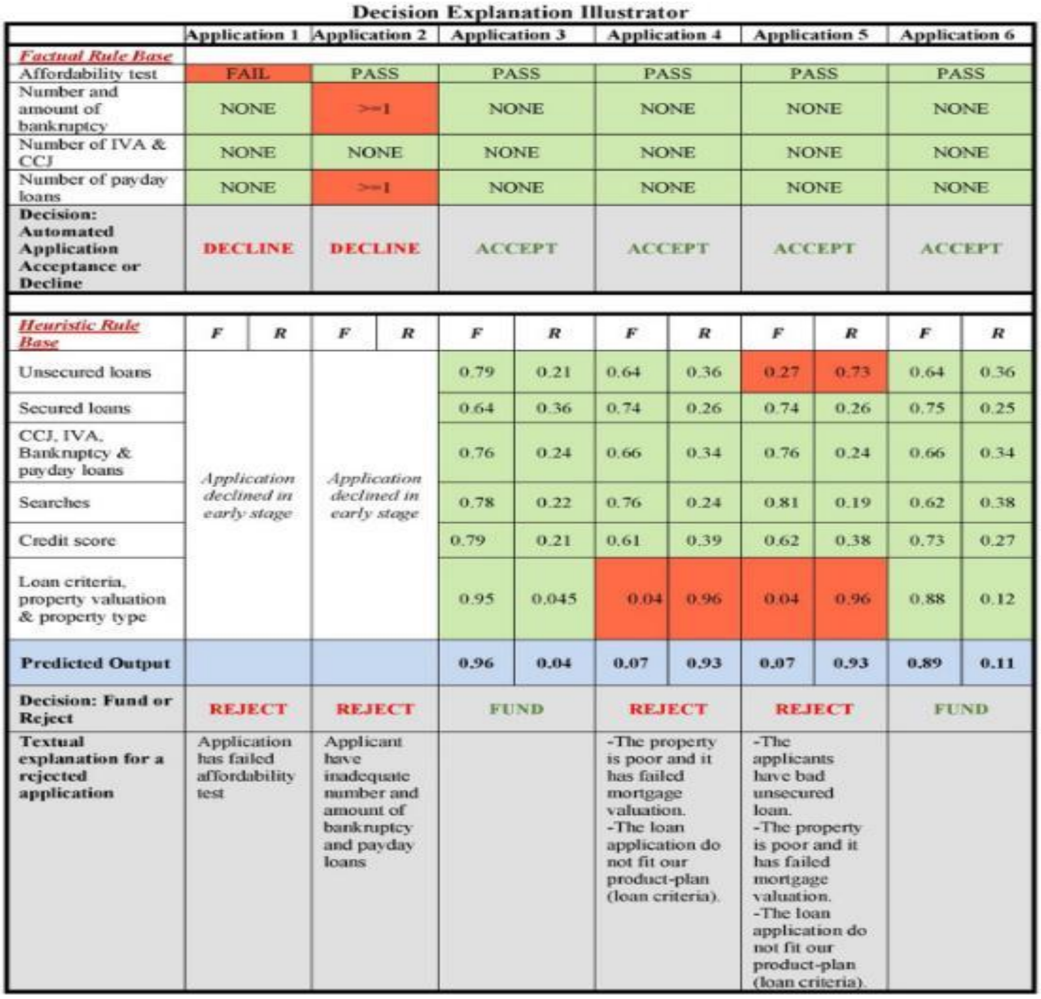
ADDT B : Addtech AB återköp av

Figure 3: Monitoring stock market activity [https://www.trapets.com/services/instantwatch-market/]

Collating features into charts and tables provides the analyst with a summary that can be interpreted in terms of rules. Indeed, an experienced analyst might recognise recurring 'patterns' across different instances. That is, similar activities might produce displays that have similar visual appearance that the analyst can associate with particular 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, 1989) it does not provide access to the underling rules the AI system used to generate the clusters (and could result in the user either anthropomorphising these rules or making assumption about which rules could have been applied).

Beliefs

In figure 4, rules used to reach a loa decision are listed, together with an indication of whether the rules have been met or breached (with pass / fail, colour coding, accept / decline). In this way, the computer's rules are exposed to the human decision maker. The textual explanation, at the bottom of figure 4, is clear and concise. What is not apparent here is whether the user is able to apply counter-factuals to the decision. For example, if we consider the column for 'Application 4' in figure 4, the heuristic rule base identifies 'loan criteria, etc.' as below criteria, but what might happen if the applicant was able to amend this?



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 4. Illustrating Beliefs in loan underwriting [Sachan et al., 2020]

Policy

In Deep (or Reinforcement) Learning, the AI system seeks to discover a Policy by which it can optimize reward (say, success in play a game) by performing Actions in specific situations. Accounts which reflect specific policy (in terms of the actions that AI systems take in response to situations) can be created as saliency maps (Greydanus et al., 2018). The saliency map can be used to infer the strategy that is being applied. While this need not reflect the policy (in terms of relationship between action and rewards that the agent is learning) it can allow the human analyst to form beliefs as to how the agent might behave in similar circumstances. However, it is not so easy to discern why the features were defined as Relevant, or even whether the AI system 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 5). But this puts the onus on the user to infer an 'explanation' of the AI system's decision-making.

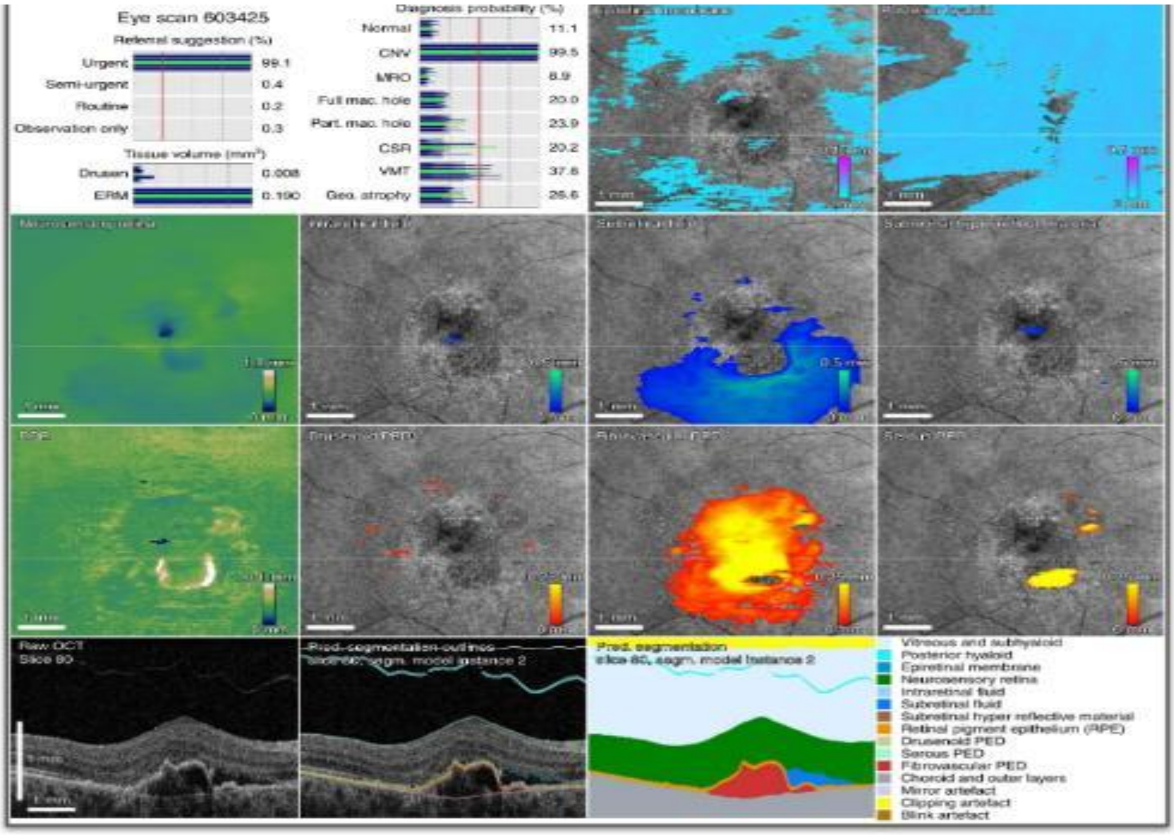


Figure 5. Multiple views of deep learning outputs for retinal diseases diagnosis [De Fauw et al., 2018]

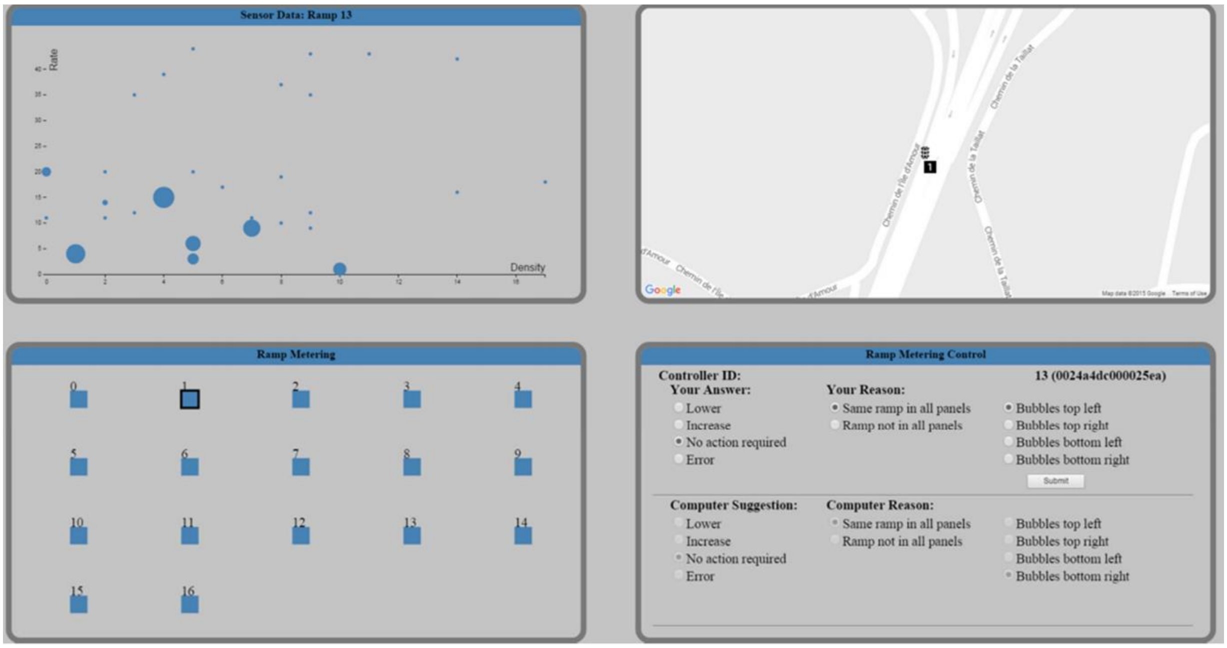
Explanatory Discourse

In robotics, humans can ask questions of the robot that allow it to explain its reasoning. Fox et al's. (2017) eXplAInable Planning (XAIP) requires the robot to justify why it chose particular actions, etc. In terms of our notion of explanatory discourse (discussed earlier) this supports alignment of situation model (in terms of the features which are attended) or alignment of relevance (in terms of justification for an action). Recognising that human explanations can make use of choices of action in a situation, Borgo et al.'s (2018) developed XAI-PLAN. While the motivation for this, and similar work, seems to be the assumption that human explanations involve evidence and counter examples, it is not apparent that the ability to answer 'why' questions necessarily involves generation of a counter example. In other words, the issue of how the situation model or definition of relevance can be challenged has received less attention to date.

EXPLANATION AND DISAGREEMENT

In an experiment in which human and computer cooperate on an actions in a road traffic management task (Morar and Baber, 2017), the human participant needs to choose an action to manage a road network (defined by traffic volume and flow) and location on the map (figure 6). A computer provides suggestions as to which action to perform. Sometimes the computer is wrong. In this experiment, the Situation involves monitoring road traffic to reduce congestion (by altering traffic rate, through control using traffic lights, to reduce density). The user is provided with information on the identification of a ramp to join or leave a highway (shown as the highlighted box in the 'ramp metering' window), the location of the ramp (shown on the map), and the state of the road network at that location (shown by the bubble chart of traffic density and rate). To read the bubble chart, the following heuristics are applicable: Low density, Low rate: no response; High density, High rate: reduce rate; High density, Low rate: increase rate; Low density, High rate: no response.

In the 'ramp metering control' window (bottom right, figure 6), the user indicates which action to take, a reason for this action, and their interpretation of the bubble chart. Below this, the computer provides its solution. Users compare their responses with that of the computer, decide whether they wish to alter their response, and then use the 'submit' button to confirm this response.



Controller ID: 13 (0024a4dc000025ea)

Your Answer:
○ Lower
○ Increase
● No action required
○ Error

Your Reason:
● Same ramp in all panels
○ Ramp not in all panels

● Bubbles top left
○ Bubbles top right
○ Bubbles bottom left
○ Bubbles bottom right

Computer Suggestion:
○ Lower
○ Increase
● No action required
○ Error

Computer Reason:
● Same ramp in all panels
○ Ramp not in all panels

○ Bubbles top left
○ Bubbles top right
○ Bubbles bottom left
● Bubbles bottom right

Figure 6: User Interface for Road Traffic Task

Table 1 translates the requirements for this experiment into the terminology of our process model, to illustrate how concepts of relevance apply to different aspects of the experiment and how these might be aligned (by the human) in their decision making.

Table 1. *Mapping the Road Traffic Decision Task to the Process Model*

Element	Situation Model	Relevance	Alignment
$S_{x1} \approx S_{x2}$	Traffic on map; Behaviour of traffic on graph; ramp being monitored	Feature: ramp id / location Cluster: bubbles on graph Belief: definition of congestion	agree Features
$R_{x1} \neq R_{x2}$	Radio-buttons selected in Ramp Metering Control panel	Belief: Mismatch between Computer and User answers	agree Belief
$\Delta R_{x2} \approx r_{x1} \sqsubseteq R_{x1}$	Selection of radio-buttons in situation model should agree	Belief: Computer is correct	agree Belief
$A_2 = \Delta s_2$	User Acts to change the Situation	Policy: Press ‘Submit’ button’	User is responsible for decision

An observation from Morar and Baber (2017) is that, rather than the computer supporting the users’ decisions, there was often an assumption that the ‘computer’ provided information requiring confirmation. This is especially problematic if ‘computer’ reliability is “quite high”, i.e., >80%, because it requires scrutiny of the recommendation at a level of detail that is not required if reliability is ‘low’ or ‘perfect’. The Situation model uses Features attended by human and computer and a common user error was to miss differences relating to ramp id. In terms of Relevance there was a need to align Beliefs about the Situation, and a common user error was not to recognise computer misinterpretation of the graph, especially when this compounded ramp id errors.

CONCLUSION

By way of conclusions, we offer some basic guidelines for the implementation of XAI:

1. Explanation should be related to beliefs about the relationship between features that can directly affect the situation being explained (situation model), or can explain the majority of the situation (explanatory power), and are plausible (construct validity);
2. The Explanation should relate the goals of the explainer and explainee.
3. The explanation to suit the explainee’s definition of relevance;
4. Explanations should be interactive and involve the explainee in the explanation;
5. Explanations should be (where necessary) actionable. The explainee should be given information that can be used to perform and/or improve future actions and behaviours;
6. There should be clarity in the definition of relevance used in the explanation: Define clusters (i.e., statistical model), belief (i.e., causal model) and policy (i.e., implications for action);
7. Explanatory discourse should allow challenge and the use of counter-examples to test the situation models and definitions of relevance that are employed in the explanation.

ACKNOWLEDGMENTS

The ideas developed in this paper have arisen from a study on ESRC Centre for Research and Evidence on Security Threats: CREST Project 277: Human Engagement Through Artificial / Augmented Intelligence.

REFERENCES

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

Baber, C., Conway, G., Attfield, S., Rooney, C., Kodagoda, N., and Waller, R. (2015) How military intelligence personnel collaborate on a sense-making exercise, *12th International Naturalistic Decision Making Conference*, McLean, VA.

Baber, C., Attfield, S., Conway, G., Rooney, C. and Kodagoda, N. (2016) Collaborative sensemaking during Intelligence Analysis exercises, *International Journal of Human Computer Studies*, *86*, 94-108.

Baber, C., McCormick, E. and Apperly, I. (2021) A framework for explainable AI, *Contemporary Ergonomics 2021.*

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

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

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

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

Greydanus, S., Koul, A., Dodge, J. and Fern, A. (2018) Visualizing and understanding Atari agents, *arXiv*:1711.00138v5

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

Hoffman, R., Miller, T., Mueller, S. T., Klein, G., and 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

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H. and Wortman Vaughan, J. (2020) Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-14.

Klein, G.A. (1989) Recognition-primed decisions, In W. B. Rouse (ed.) *Advances in Man-Machine Systems Research*, *volume 5,* Greenwich, CT: JAI Press, 47-92.

Klein, G., Phillips, J.K., Rall, E.L. and Peluso, D.A. (2007) A data-frame theory of sensemaking, *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*, New York, NY: Lawrence Erlbaum Assoc Inc., 113-155.

Langley, P., (2019) Varieties of Explainable Agency, *ICAPS Workshop on Explainable AI Planning*.

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.

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

Miller, T. (2017) Explanation in Artificial Intelligence: insights from the Social Sciences, *arXiv*:1706.07269.

Miller, T., Howe, P., and Sonenberg, L. (2017) Explainable AI: beware of inmates running the asylum, *Proceedings IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*

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. and van Diggelen, J. (2018) Using perceptual and cognitive explanations for enhanced human-agent team performance, In Harris, D. (ed.) *EPCE 2018*, LNCS (LNAI), 10906, Springer, Cham, 204–214.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135-1144.

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

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*.

Sperber, D. and Wilson, D. (1986) *Relevance: Communication and Cognition*, Cambridge, MA: Harvard University Press.

Springer, A. (2019) Enabling effective transparency: Towards user-centric intelligent systems, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 543–544.