A framework for explainable AI

Chris Baber[1], Emily McCormick[1] & Ian Apperley[2]

[1]School of Computer Science, University of Birmingham, [2]School of Psychology, University of Birmingham

ABSTRACT

The issue of 'explanation' has become prominent in automated decision aiding, particularly when those aids rely on Artificial Intelligence (AI). In this paper, we propose a formal framework of 'explanation' which allows us to define different types of explanation. We provide a use-cases to illustrate how explanation can differ, both in human-human and human-agent interactions. At the heart of our framework is the notion that explanation involves common ground in which two parties are able to align the features to which they attend and the type of relevance that they apply to these features. Managing alignment of features is, for the most part, relatively easy and, in human-human explanation, people might begin an explanation by itemizing the features they are using (and people typically only mention one or two features). However, providing features without an indication of Relevance is unlikely to provide a satisfactory. This implies that explanations that only present features (or Clusters of features) are incomplete. However, most Explainable AI provides output *only* at the level of Features or Clusters. From this, the user has to infer Relevance by making assumptions as to the *beliefs* that could have led to that output. But, as the reasoning applied by the human is likely to differ from that of the AI system, such inference is not guaranteed to be an accurate reflection of *how* the AI system reached its decision. To this end, more work is required to allow interactive explanation to be developed (so that the human is able to define and test the inferences and compare these with the AI system's reasoning).

KEYWORDS

Explainable AI, Human-Agent Interaction

Introduction

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.…"*[1] Providing an 'explanation' for the output of an AI system *ought* to make it easier for a human to understand the output. One approach to AI explanation is to focus on the algorithm used by the AI system: *"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). Such understanding could allow the human to challenge the output by either offering information that the computer had not considered or provide the computer with counter-factual examples. Often, the ultimate purpose of AI explanation is to allow the human to accept responsibility for the consequences arising from this output; be it in the form of medical diagnosis or treatment recommendation, or decisions on loan applications, or

[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.

actions made by autonomous vehicles. However, creating 'explainable AI' is not trivial (Arrieta et al., 2019; Mueller et al., 2019), and a significant part of this problem lies in developing a clear definition of 'explanation'. For an Ergonomics perspective, an 'explanation' ought not to be solely the concern of the 'user' of an AI system. But this should concern all humans who communicate *through* the AI systems, i.e., the humans who programme and deploy the AI systems, the analysts who use AI systems, people who collect data that will be used by the AI system, and the managers and other stakeholders who use the output from the AI systems. From this, we claim that there is not one type of 'explanation' but several. We present a framework in which these different types of explanation can be realized. We are interested in the boundaries across which agents (human or AI) will share information and the nature of 'explanation' required across these boundaries.

Defining explanation

In an early attempt at a formal definition of explanation, Hempel (1924) proposed a 'Covering Law Model' of History. A core question for historians is *why* a given Event, *E*, occurred. Hempel suggested that a set of prior events (or states) could be regarded as antecedent Causes, and these are combined according to some 'Law' that the historian proposed. From this, an argument could be presented (either deductively or inductively) that the occurrence of antecedents increases the probability of E occurring. While the term 'Law' might feel overly reductive, Hempel's (1924) approach offers us an opportunity to define an Explanation in terms of specific elements and their relations to each other.

Elements of Explanation

Our aim is to produce a formal description that can reflect different types of explanation, that is applicable to human-human conversation and human-AI interaction, and that allows us to ask *how* explanations are produced. This is represented in a framework (figure 1).

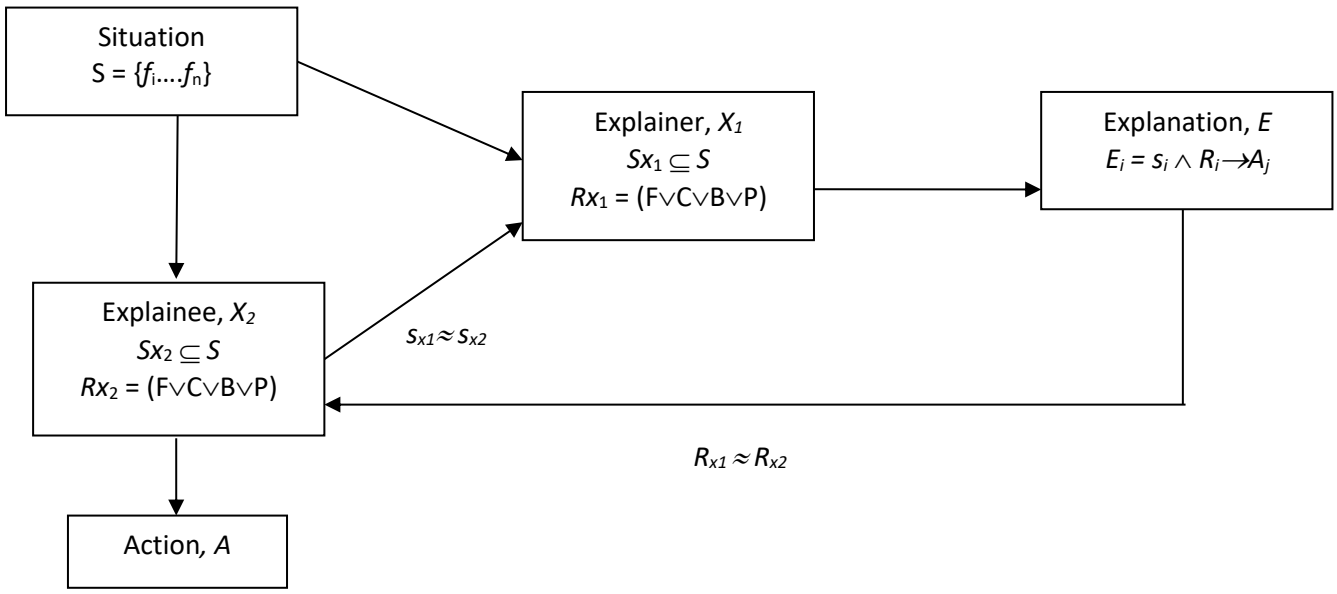

Figure 1: A framework for explanation

From [1], an Explanation is generated when two parties, X_1 and X_2, in a Situation, S, seek to align the sets of features of the Situation 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. Actions could be, for

example, that X_2 acknowledges or accepts the Explanation, that X_2 challenges the explanation or seeks further information, that X_2 performs some task as a result of the explanation.

Elaborating on this, an explanation, E, occurs in, and relates to, a situation. A situation, S, has a set of features, $\{f_i \ldots f_n\}$, which can be described symbolically, using words, numbers, pictures, etc. A 'feature' is some aspect of the situation to which people can attend and individuals in a situation ground their Situation Awareness, s_i, by attending to a subset of all features in S, i.e., $s_i \subseteq S$. Features are external to individuals, in that anyone in S ought to be able to attend to the same features. However, we accept that there will be situations in which some features might either in internal (i.e., known by an individual) or not be immediately accessible to all parties.

The first challenge in producing an explanation is to ensure that the set of features to which the Explainer, X_1, attends will overlap with the set used by the Explainee, X_2. This means that there should be approximate equivalence between these sets of features, i.e., $s_{x1} \approx s_{x2}$. Notice that we do assume that these sets are *identical*, only for there to be sufficient overlap, i.e., 'common ground' (Clark, 2015). 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, and so there is an obvious need to share features which contribute to the interpretation of the situation.

From this, a second challenge is to agree on which features define the Situation, i.e., to define 'relevance'. Relevance, R, can be defined in terms of four levels:

- *Features, F:* features in the situation to which both parties can attend;
- *Clusters, C*: features which typically co-occur in similar situations;
- *Beliefs, B*: the reason why clusters co-occur, and which can predict consequences if specific features alter;
- *Policies, P*: rules which allow actions to be linked to clusters or features.

Thus, Relevance is defined in terms of F or C or B or P, i.e., $Rx_i = (F \vee C \vee B \vee P)$.

From this, we propose that an Explanation, *E*, involves the set of Feature, $\{f_i \ldots f_n\}$, to which a person attends in a situation, S, a means of defining the relevance, *R*, in terms of F, C, B, or P, of these features and a (potential) aim of influencing Action, *A*:

$$E_i = s_i \wedge R_i \rightarrow A_j \quad , \text{ where } R = (F \vee C \vee B \vee P) \qquad [1]$$

An Explanation 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 1 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.

Scenario

Having defined elements of Explanation, we apply these to use-cases to illustrate the processes that might occur:

> *A hacker has obtained access to email accounts in your organization 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 have to create new passwords.*

In this Scenario, the *Situation* has *features* that include measures of network activity, e.g., messages across a network constitute a *Feature*, a count of messages over time constitutes a *Cluster*, whether the network is 'normal' or 'unusual' a *Belief*, and responses to manage the network is a *Policy*.

Examples of Explanation: human-human

Example H-H.I: $S_{x1} \approx S_{x2}$ *and* $R_{x1} \approx R_{x2}$*:* In this instance, both parties assume that $S_{x1} \approx S_{x2}$ and $R_{x1} \approx R_{x2}$, and the need for explanation is negligible. However, when $S_{x1} \neq S_{x2}$ the individuals will need to resolve common ground, e.g., agree which features are relevant. If Explainer, X_1, and Explainee, X_2, have similar knowledge, training, experience etc., i.e., $X_1 \approx X_2$, then alignment could involve indicating a change in a relevant feature. We assume there is 'honest signaling' (Maynard Smith and Harper, 2003) in that the feature is relevant to the situation. For example, the email traffic in the network might be unusually low for a Tuesday. In this case, X_1 might draw the attention of X_2 to this feature. However, if X_2 does not recognize the relevance of this feature, then an explanation would involve X_1 both highlighting the feature and presenting the Belief as to its relevance. Here, (because $X_1 \approx X_2$ is the equilibrium state) X_2 should interpret the Belief with minimal effort, i.e., X_1 can highlight the relevant features and expect X_2 to access a Belief to determine relevance.

Example H-H.II: $S_{x1} \approx S_{x2}$ *and* $R_{x1} \neq R_{x2}$ *:* For people without similar backgrounds (i.e., $X_1 \neq X_2$) alignment will be more effortful. As an initial move, the focus on features would allow people to check their assumption that alignment is possible, or the Explainer could encourage the Explainee to infer an appropriate Feature, Cluster, Belief or Policy. A characteristic of explanation, particularly in social-cognitive psychology, is that people are likely to offer one or two features as first-pass explanation (McClure et al., 2001; Leddo et. al., 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) to be sufficient to explain the situation.

Example H-H.III: $S_{x1} \approx S_{x2}$ *and* $R_{x1} \neq R_{x2}$ *and* $\Delta R_{x2} \approx r_{x1} \subseteq R_{x1}$*:* 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 alignment (that is, the apprentice will not 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, i.e., $\Delta R_{x2} \approx r_{x1} \subseteq R_{x1}$. In order for this to occur, there is a need to establish that $S_{x1} = S_{x2}$. In this case, an explanation (a) ensures that X_2 attends to specific features, in order to (b) encourages 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.

Example H-H.IV: $S_1 \neq S_2$ and $R_1 \neq R_2$ and $\Delta R_2 \approx r_1 \subseteq R_1$ and $A_2 = \Delta s_2$: While Example III emphasizes explanation as an epistemic objective (to increase knowledge of X_2), 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), i.e., $A_{x2} = \Delta s_{x2}$. In other words, the emphasis is on motivating the user to change a password, etc. It is arguable whether this motivational objective is fully dependent on a change in knowledge, e.g., does it matter if the user does not understand the entire basis of the advice as long as they act as required? In this case, the explanation places 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 X_2 to have a 'productive understanding'. Indeed, an aim would be for X_2 to become their own Explainer (or to have the analyst-as-explainer replaced by another source, such as a leaflet, web-site, etc.).

Example H-H.V: $S_{x1} \neq S_{x2}$ and $R_{x1} \neq R_{x2}$: Assume that 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. Further, assume that the newspaper reader is not an IT specialist. 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.

Example H.H.VI: $S_{x1} \neq S_{x2}$ and $R_{x1} \approx R_{x2}$: Assume that a formal report is written following the incident. This report is consulted by other analysts (possibly in other organizations). 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. So, 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, becoming more focused (Kempe et al., 2019). This might be the result of the formal structures imposed by the style of reports; it might be the result of the manner in which people share information; or it might be the result of a desire to focus on relevant information. A consequence of this might be that fewer of the Features of the original situation 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 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 by the report 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) refers to this as 'explanation evaluation' and concludes that important criteria to evaluate an explanation are: probability, simplicity, generality, and coherence with prior beliefs. So, in our hacking example, X_2 is most likely to accept an explanation that a) is consistent with their beliefs about email hacking (coherence); b) includes fewer causes but can be related to events they have experienced (simplicity, generality); and c) that a particular type of attack is a '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, 2005). It is also worth noting that whilst 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).

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. Before exploring these, it is worth noting that 'explanation' is not always something that humans find easy to perform. This could be because we are not able to reflect on the features we are using, or to articulate the concept of relevance that we apply, or we might not be able to take the perspective of our Explainee and so cannot determine their beliefs. If humans struggle with 'explanation' we should not be surprised to find that this is a challenge for AI systems.

Example H.A.I: $S_1 \approx S_2$ *and* $R_1 \approx R_2$*:* Recommender Systems might inform their users of the specific features that inform the recommendation, e.g., a word-cloud taken from movie reviews (Gedkili et al., 2014) or a histogram of ratings of a movie by 'similar' users (Herlocker et al., 2000). Here, Relevance is presented as a Cluster. Alternatively, *ExpertClerk* (Shimazu, 2002) offers 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." Objects are compared in terms of features and the trade-off is presented as a Belief.

Example H.A.II: $S_{x1} \approx S_{x2}$ *and* $R_{x1} \neq R_{x2}$*:* In most applications of Machine Learning (ML), identifying a Cluster does *not* involve a Belief[2]. In this case, while there might be an intention of aligning the features that ML algorithms use with those that the human can interpret, such that $S_1 \approx S_2$, it is much more difficult to align Relevance. However, users might *assume* Belief from the output of ML, e.g., either anthropomorphizing 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. Association-Rule Mining, for example, can be used to highlight dependencies between features that are more akin to our notion of 'belief' (Altaf et al., 2017).

Example H.A.III: $S_{x1} \approx S_{x2}$ *and* $R_{x1} \neq R_{x2}$ *and* $\Delta R_{x2} \approx r_{x1} \subseteq R_{x1}$: Educational technologies provide personalized and adaptive environments to support learning (Dawson et al., 2010). In these systems, learners are provided with situations in which they review material (features) in order to answer questions (action), and performance on the questions will impact on progression their 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.

Example H.A.IV: $S_1 \neq S_2$ *and* $R_1 \neq R_2$ *and* $\Delta R_2 \approx r_1 \subseteq R_1$ *and* $A_2 = \Delta s_2$: Technology-mediated 'nudging' (Caraban et al., 2019) creates '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 well-being. These technologies remind the user of consequences of their actions, suggest alternative actions, or

[2] We note that the word 'belief' is used in some forms of Machine Learning, but has quite a different meaning to the way we use it. For instance, in a Bayesian Belief Network (BBN) situation features are arranged in a network. 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. 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, generalizable 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.

emphasize social desirability of the consequences. In our terms, the focus is on Action, through highlighting relevant Beliefs.

Example H.A.V: $S_{x1} \neq S_{x2}$ *and* $R_{x1} \neq R_{x2}$*:* 'Explainable AI' (Arrieta et al., 2020) tends to involve a Situation in which the explainer is the AI system, which attends to a set of features in terms of a Policy. In Deep (or Reinforcement) Learning, the AI Policy will optimize 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, Greydanus et al., 2017) 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 easy to discern how the Features were defined as Relevant.

Example H.A.VI: $S_{x1} \neq S_{x2}$ *and* $R_{x1} \approx R_{x2}$*:* 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 emphasize or notion of relevance. User interfaces for argumentation visualize 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).

Discussion

A framework is developed 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 agreement on the features (in data sets or a situation) to which explainer and explainee attend, and agreement as to why these features are relevant (and we propose three levels of relevance, i.e., 'cluster' in which a group of features will 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 on relevance) depends on the knowledge and experience of explainer and explainee, and much of the process of explanation involves ensuring alignment in terms of knowledge and experience. Thus, 'Explanation' is the process by which common ground is established and maintained. From our framework of explanation, we propose the following guidelines:

- Explanations should highlight **Relevance**: include the relationship between features of a situation and the event being explained, and should be plausible in terms of a concept of Relevance agreed between Explainer and Explainee.
- Explanations should include relevant **Features:** Explainer and Explainee should agree key features of the situation.
- Explanations should be Framed to suit the **audience:** fit the explanation to the explainee's understanding of the situation and goals.
- Explanations should be **interactive:** involve the explainee in the explanation.
- Explanations should be (where appropriate) **actionable:** the explainee should be given information that can be used to perform and improve future actions and behaviours.

Acknowledgement

The work reported in this paper was supported by a grant from the ESRC / CREST [Project: 277].

References

Altaf, W., Shahbaz, M. and Guergachi, A. (2017) Applications of association rule mining in health informatics: a survey. *Artificial Intelligence Review*, *47*, 313-340.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R. (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, *58*, 82-115.

Bartlett, F. C. (1932) *Remembering.* Oxford, Macmillan.

Caraban, A., Karapanos, E., Gonçalves, D. and Campos, P. (2019) 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction, *2019 CHI*, 1-15.

Dawson, S., Heathcote, L., and Poole, G. (2010) Harnessing ICT potential: The adoption and analysis of ICT systems for enhancing the student learning experience, *International Journal of Educational Management, 24*, 116-128

Gedikli, F., Jannach, D. and Ge, M. (2014) How should I explain? A comparison of different explanation types for recommender systems, *International Journal of HumanComputer Studies*, *72*, 367–382.

Greydanus, S., Koul, A., Dodge, J. and Fern, A., 2017, Visualizing and understanding atari agents. *arXiv preprint arXiv:1711.00138*.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S. and Turini, F. (2019) Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*.

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*, 843-887.

Hempel, C.G. (1924) The function of general laws in history, *Journal of Philosophy, 39*, 35-48

Herlocker, J.L., Konstan, J.A. and Riedl, J. (2000) Explaining collaborative filtering recommendations, *ACM conference on Computer Supported Cooperative Work*, 241–250.

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

Kempe, V., Gauvrit, N., Gibson, A., & Jamieson, M. (2019) Adults are more efficient in creating and transmitting novel signalling systems than children, *Journal of Language Evolution*, *4*, 44-70.

Klein, G., Phillips, J.K., Rall, E.L. and Peluso, D.A. (2007) A data–frame theory of sensemaking. In *Expertise out of context*, Psychology Press, 118-160.

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*, 933.

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

McClure, J., Hilton, D., Cowan, J., Ishida, L., & Wilson, M. (2001) When rich or poor people buy expensive objects: Is the question how or why. *Journal of Language and Social Psychology*, *20*, 229-257.

Maynard Smith, J. and Harper, D., 2003, *Animal Signals*, Oxford: Oxford University Press.

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

Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A. and Klein, G., (2019) Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*.

Reed, C., Budzynska, K., Duthie, R., Janier, M., Konat, B., Lawrence, J., Pease, A. and Snaith, M. (2017) The argument web: An online ecosystem of tools, systems and services for argumentation. *Philosophy & Technology*, *30*, 137-160.

Shimazu, H. (2002) ExpertClerk: A Conversational Case-Based Reasoning Tool for Developing Salesclerk Agents in E-Commerce Webshops, *Artificial Intelligence Review*, *18*, 223-244.

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