# Explainable AI: Beware of Inmates Running the Asylum

**Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences**

**Tim Miller**[*] and **Piers Howe**[†] and **Liz Sonenberg**[*]
[*]School of Computing and Information Systems
[†]Melbourne School of Psychological Sciences
University of Melbourne, Australia
{tmiller,pdhowe,l.sonenberg}@unimelb.edu.au

## Abstract

In his seminal book *The Inmates are Running the Asylum: Why High-Tech Products Drive Us Crazy And How To Restore The Sanity* [2004, Sams Indianapolis, IN, USA], Alan Cooper argues that a major reason why software is often poorly designed (from a user perspective) is that programmers are in charge of design decisions, rather than interaction designers. As a result, programmers design software for themselves, rather than for their target audience; a phenomenon he refers to as the '*inmates running the asylum*'. This paper argues that explainable AI risks a similar fate. While the reemergence of explainable AI is positive, this paper argues most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users. But explainable AI is more likely to succeed if researchers and practitioners understand, adopt, implement, and improve models from the vast and valuable bodies of research in philosophy, psychology, and cognitive science; and if evaluation of these models is focused more on people than on technology. From a light scan of literature, we demonstrate that there is considerable scope to infuse more results from the social and behavioural sciences into explainable AI, and present some key results from these fields that are relevant to explainable AI.

## 1 Introduction

> "*Causal explanation is first and foremost a form of social interaction. One speaks of giving causal explanations, but not attributions, perceptions, comprehensions, categorizations, or memories. The verb to explain is a three-place predicate:* Someone *explains* something *to* someone. *Causal explanation takes the form of conversation and is thus subject to the rules of conversation.*" — Hilton [1990].

The term "explainable AI" has regained traction again recently, after being considered important in the 80s and 90s in expert systems particularly; see [Chandrasekaran *et al.*, 1989], [Swartout and Moore, 1993], and [Buchanan and Shortliffe, 1984]. High visibility of the term, sometimes abbreviated XAI, is seen in grant solicitations [DARPA, 2016] and in the popular press [Nott, 2017]. One area of explainable AI receiving attention is explicit *explanation*, on which we say more below.

While the title of the paper is deliberately tongue-in-cheek, the parallels with Cooper [2004] are real: leaving decisions about what constitutes a good explanation of complex decision-making models to the experts who understand these models the best is likely to result in failure in many cases. Instead, models should be built on an understanding of explanation, and should be evaluated using data from human behavioural studies.

In Section 2, we describe a simple scan of the 23 articles posted as 'Related Work' on the workshop web page. We looked at two attributes: whether the papers were built on research from philosophy, psychology, cognitive science, or human factors; and whether the reported evaluations involved human behavioural studies. The outcome of this scan supports the hypothesis that ideas from social sciences and human factors are not sufficiently visible in the field.

In Section 3, we present some key bodies of work on explanation and related topics from social and behavioural sciences that will be of interest to those in explainable AI, and briefly discuss what their impact could be.

## 2 Explainable AI Survey

To gather some data to test the hypothesis that the social sciences and human behavioural studies are not having enough impact in explainable AI, a short literature survey was undertaken. This survey is not intended to be even close to comprehensive – it is merely illustrative. However, the results that it shows are reflective of many other papers in the area that the authors have read.

### 2.1 Selected Papers

The articles surveyed were taken from the 'Related Work' list that was posted on the website for the IJ-

CAI 2017 Explainable AI workshop[1] as of 16 May 2017 — the workshop to which this paper is submitted. In total 23 articles were on the list, although one was not included in the results as described later. This list can be found in Appendix A.

As noted already, this list is far from comprehensive, however, it is a useful list for two reasons:

1. First, it was compiled by the explainable AI community: the organisers of the conference requested that people send related papers to be added to the list. As such, it represents at least a subset of what the community see at the moment as highly relevant papers for researchers in explainable AI.
2. Second, it is objective from the perspective of the authors of this paper. We did not contribute to the list, so the selection is not biased by our argument.

While the authors of some of the listed papers may not consider their work as explainable AI, almost all of the papers were describing methods for automatically generating explanations of some type.

The paper that was excluded is Tania Lombrozo's survey paper on explanation research in cognitive science [Lombrozo, 2012]. This is not an explainable AI paper – indeed, it summarises one of the bodies of work of which we argue people should be more aware.

## 2.2 Survey Method

The survey was lightweight: it only looked for evidence that the presented research was somehow influenced by a scientific understanding of explanations, and that the evaluations were performed using data derived from human behaviour studies or similar. We categorised the papers on the three items of interest, with the criteria for the scores as follows:

1. *On topic*: Each paper was categorised as either being about explainable AI or not, based on our understanding of the topic. It is possible that some papers were included on the workshop website because they presented good challenges or potentially useful approaches, but were not papers about explanation *per se*, in which case they were 'off topic'.
2. *Data Driven*: Each paper was given a score from 0–2 inclusive.

   A score of 1 was given if and only if one or more of the *references* of the paper was an article on explanation in social science, meaning that: (a) explanation or causal attribution as done by humans is one of the main topics of the referenced article(s); (b) the referenced article(s) validated their claims using data collected from human behaviour experiments; and (c) the referenced article(s) appear in a non-computer science venue *or* in a computer science venue but contributed to the understanding of explanation in general (outside of AI).

   A score of 2 was given if and only if (a), (b), and (c) above held, and the survey article (not the referenced article) described an algorithm for automatically generating explanations and this algorithm was derived from data from the social sciences. In other words, the algorithm is explicitly based on a model from one or more of the references.

   A score of 0 was given for any other paper; that is, no references satisfying (a), (b), and (c).
3. *Validation*: Each paper was given a binary 0/1. A score of 1 was given if and only if the evaluation in the survey article (note, not the referenced article) was based on data from human behavioural studies. Even if the algorithm is categorised as data driven, we argue that it is still important to test that the assumptions and trade-offs made are suitable. It is therefore necessary to (eventually) perform behavioural studies to test if the explanations produced by the algorithm are appropriate for humans.

## 2.3 Results

Table 1 shows the results for the survey. Results for each of the surveyed articles are available in Appendix B. Five papers were deemed 'off topic', however, the results are included because we could not know the intent of those who submitted articles to the reading list. For the 'Data driven' entry, column 'N' means that we were unsure about the reference. In this case, one paper had a reference to a cognitive science article of which we were unable to locate a copy. For the 'Validation' entry, column 'N' means 'not applicable': three papers were categorised as not applicable because their status were not research articles, but review articles or position papers, and thus, they did not present any algorithm or model to evaluate.

| | **On topic** (17 articles) | | | | **Off topic** (5 articles) | | | |
|---|---|---|---|---|---|---|---|---|
| **Criterion** | N | 0 | 1 | 2 | N | 0 | 1 | 2 |
| Data driven | 1 | 11 | 4 | 1 | 0 | 4 | 0 | 1 |
| Validation | 3 | 10 | 4 | — | 0 | 4 | 1 | — |

Table 1: Results on small survey

These results show that for the on-topic papers, only four articles referenced relevant social science research, and only one of them truly built a model on this. Further, serious human behavioural experiments are not currently being undertaken. For off topic papers, the results are similar: limited input from social sciences and limited human behavioural experiments.

## 2.4 Discussion

The results, while only on a small set of papers, provide evidence that many models being used in explainable

[1] See `http://home.earthlink.net/~dwaha/research/meetings/ijcai17-xai/`.

AI research are not building on current scientific understanding of explanation. Further, human behavioural experiments are rare — something that needs to change for us to produce useful explanatory agents.

It is important to note that we are not interpreting the above observations to say that there is not a lot of excellent research on explainable AI. For example, consider Ribeiro *et al.* [2016], who have done some remarkable work on explaining classifiers, and yet scored '0' on the 'Data Driven' criteria. Instead, they have constructed their own understanding of how people evaluate explanations for their particular field over a series of human behavioural experiments. However, developing such an understanding will not always be required or even possible for many researchers, so in these cases, building on social science research is a sound place to start.

# 3 Where to? A Brief Pointer to Relevant Work

In the different sub-fields of social sciences, there are several hundred articles on explanation, not to mention another entire field on causality. It is not feasible to expect that AI researchers and practitioners can navigate this entire field in addition to their own field of expertise, especially considering that the relevant literature is written for a different audience. However, there are some key areas that should be of interest to those in explainable AI, which we outline in this section.

Miller [2017] provides an in-depth survey of all articles cited in this section plus many other relevant articles, and draws parallels between this work and explainable AI. Here, we present several key ideas from that work to demonstrate ways that models of explainable AI can benefit from models of human explanation.

## 3.1 Contrastive Explanation

Perhaps the most important result from this work is that explanations are *contrastive*; or more accurately, *why–questions* are contrastive. That is, why–questions are of the form "*Why* P *rather than* Q*?*", where $P$ is the *fact* that requires explanation, and $Q$ is some *foil* case that was expected. Most philosophers, psychologists, and cognitive scientists in this field assert that *all* why–questions are contrastive (e.g. see [Hilton, 1990; Lombrozo, 2012; Miller, 2017]), and that when people ask for an explanation "*Why* P*?*", there is an implicit contrast case. Importantly, the contrast case helps to frame the possible answers and make them relevant [Hilton, 1990]. For example, explaining "*Why did Mr. Jones open the window?*" with the response "*Because he was hot*" is not useful if the implied foil is Mr. Jones turning on the air conditioner, as this explains both the fact and the foil; or if the implied foil was why Ms. Smith, who was sitting closer to the window, did not open it instead, as the cited cause does not refer to a cause of Ms. Smith's lack of action.

This is a challenge for explainable AI, because it may not be easy to elicit a contrast case from an observer. However, it is also an opportunity: as Lipton [1990] argues, answering a contrastive question is often easier than giving a full cause attribution because one only needs to understand the difference between the two cases, so one can provide a complete explanation without determining or even knowing all causes of the event.

## 3.2 Attribution Theory

Attribution theory is the study of how people attribute causes to events; something that is necessary to provide explanations. It is common to divide the types of attribution into two classes: (1) causal attribution of social behaviour (called *social attribution*); and (2) general causal attribution.

**Social Attribution** The book from Malle [2004], based on a large body of work from himself and other researchers in the field, describes a mature model of how people explain behaviour of others using folk psychology. He argues that people attribute behaviour based on the beliefs, desires, intentions, and traits of people, and presents theories for why failed actions are described differently than successful actions; the former often referring to some precondition that could not be satisfied.

Malle's work provides a solid foundation on which to build social attribution and explainable AI models for many sub-fields of artificial intelligence. Social attribution is important for systems in which *intentional action* will be cited as a cause; in particular, it is important for systems doing deliberative reasoning, and the concepts used in his work are closely linked to that of systems such as *belief-desire-intention* models [Rao and Georgeff, 1995] and AI planning.

**Causal Connection** Research on how people connect causes shows that they do so by undertaking a mental simulation of what *would have happened* had some other event turned out differently [Kahneman and Tversky, 1982; Hilton *et al.*, 2005; McCloy and Byrne, 2000].

However, simulating an entire causal chain is infeasible in most cases, so cognitive scientists and social psychologists have studied how people decide which events to 'undo' (the counterfactuals) to determine cause. For example, people tend to undo more proximal causes over more distal causes [Miller and Gunasegaram, 1990], abnormal events over normal events [Kahneman and Tversky, 1982], and events that are considered more 'controllable' [Girotto *et al.*, 1991].

For explainable AI models, these heuristics are useful from a computational perspective in large causal chains, in which causal attribution is intractable in many cases [Eiter and Lukasiewicz, 2002]. Effectively, they can be used to 'skip-over' or *discount* some events and not consider their counterfactuals, while being consistent with what an explainee would expect.

## 3.3 Explanation Selection

An important body of work is concerned with explanation *selection*. People rarely expect an explanation that consists of an actual and complete cause of an event. Instead, explainers select one or two causes and present

these as *the* explanation. Explainees are typically able to 'fill in' their own causal understanding from just these. Thus, some causes are better explanations than others: events that are 'closer' to the fact in question in the causal chain are preferred over more distal events [Miller and Gunasegaram, 1990], but people will 'trace through' closer events to more distal events if those distal events are human actions [Hilton *et al.*, 2005] or abnormal events [Hilton and Slugoski, 1986].

In AI, perhaps some models are simple enough that explanation selection would not be valuable, or visualisation would provide a powerful medium to show many causes at once. However, for causal chains with than a handful of causes, we argue that explanation selection can be used to simplify and/or prioritise explanations.

### 3.4 Explanation Evaluation

The work discussed in this section so far looks at how explainees generate and select explanations. There is also a body of work that studies how people evaluate the quality of explanations provided to them. The most important finding from this work is that the probability that the cited cause is actually true is not the most important criteria people use [Hilton, 1996]. Instead, people judge explanations based on so-called *pragmatic influences* of causes, which include criteria such as usefulness, relevance, etc. [Slugoski *et al.*, 1993].

Recent work shows that people prefer explanations that are *simpler* (cite few causes) [Lombrozo, 2007], more *general* (they explain more events) [Lombrozo, 2007], and *coherent* (consistent with prior knowledge) [Thagard, 1989]. In particular, Lombrozo [2007] shows that the people disproportionately prefer simpler explanations over more likely explanations.

These criteria are important to any work in explainable AI. Giving simpler explanations that increase the likelihood that the observer both *understands* and *accepts* the explanation may be more useful to establish trust, if this is the primary goal of the explanation. Learning from these and adding them as objective criteria to models of explainable AI is important.

### 3.5 Explanation as Conversation

Finally, it is important to remember that explanations are interactive conversations, and that people typically abide by certain rules of conversation [Hilton, 1990]. *Grice's maxims* [Grice, 1975] are the most well-known and widely accepted rules of conversation. In short, they say that in a conversation, people consider the following: (a) quality; (b) quantity; (c) relation; and (d) manner. Coarsely, these respectively mean: only say what you believe; only say as much as is necessary; only say what is relevant; and say it in a nice way. Hilton [1990] argues that as explanations are conversations, they follow these maxims. There is body of research that demonstrates people do follow these maxims, as discussed by Miller [2017].

Note that we are not arguing that explanations must be text or verbal. However, explanations presented in a visual way, for example, should have similar properties, and these maxims offer a useful set of objective criteria.

### 3.6 Where not to go

Finally, we discuss work that we believe should be discounted in explainable AI. Specifically, two well-known theories of explanation, sometimes cited and used in explainable AI articles, are the *logically deductive model* of explanation [Hempel and Oppenheim, 1948], and the *co-variation model* [Kelley, 1967]; both of which have had significant impact. However, since its publication, researchers found that the logically-deductive model was inconsistent in many ways, and instead derived new models of explanation. Similarly, the co-variation model was found be problematic and did not account for many facets of human explanation [Malle, 2011], so was refined into other models, such as those of abnormality described in Section 3.3.

While these models are still cited as part of the history of research in explanation, they are no longer considered valid models of human explanation in cognitive and social science. We contend, therefore, that explainable AI models should build on these newer models, which are widely accepted, rather than these earlier models.

## 4 Conclusions

We argued that existing models of how people generate, select, present, and evaluate explanations are highly relevant to explainable AI. Via a brief survey of articles, we provide evidence that little research on explainable AI draws on such models. Although the survey was limited, it is clear from our readings that the observation holds more generally. We pointed to a handful of key articles that we believe could be important, but for a proper presentation and discussion of these, see Miller [2017].

We encourage researchers and practitioners in explainable AI to collaborate with researchers and practitioners from the social and behavioural sciences, to inform both model design and human behavioural experiments. We do not advocate that every paper on explainable AI should be accompanied by human behavioural experiments — proxy studies are valid ways to evaluate models of explanation, especially those in early development, and computational problems are also of interest. However, we support the emphasis in the recent DARPA solicitation [DARPA, 2016] on reaching "human-in-the-loop techniques that developers can use ... for more intensive human evaluations,' and agree with Doshi-Velez and Kim [2017] that to have a real-world impact, "it is essential that we as a community respect the time and effort involved to do such evaluations."

We hope that readers of this paper and participants in the workshop agree with our position and, where feasible, adopt existing models and methods to reduce the risk that it is only the inmates that are running the asylum.

## References


[Buchanan and Shortliffe, 1984] Bruce Buchanan and Edward Shortliffe. *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, 1984.

[Chandrasekaran *et al.*, 1989] B Chandrasekaran, Michael C. Tanner, and John R. Josephson. Explaining control strategies in problem solving. *IEEE Expert*, 4(1):9–15, 1989.

[Cooper, 2004] Alan Cooper. *The inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity*. Sams, IN, USA, 2004.

[DARPA, 2016] DARPA. Explainable artificial intelligence (XAI) program. `http://www.darpa.mil/program/explainable-artificial-intelligence`, 2016. Full solicitation at `http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf`.

[Doshi-Velez and Kim, 2017] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv e-prints*, 1702.08608, 2017.

[Eiter and Lukasiewicz, 2002] Thomas Eiter and Thomas Lukasiewicz. Complexity results for structure-based causality. *Artificial Intelligence*, 142(1):53–89, 2002.

[Girotto *et al.*, 1991] Vittorio Girotto, Paolo Legrenzi, and Antonio Rizzo. Event controllability in counterfactual thinking. *Acta Psychologica*, 78(1):111–133, 1991.

[Grice, 1975] Herbert P Grice. Logic and conversation. In *Syntax and semantics 3: Speech arts*, pages 41–58. New York: Academic Press, 1975.

[Hempel and Oppenheim, 1948] Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175, 1948.

[Hilton and Slugoski, 1986] Denis J Hilton and Ben R Slugoski. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, 93(1):75, 1986.

[Hilton *et al.*, 2005] Denis J. Hilton, John L. McClure, and R. Slugoski, Ben. The course of events: Counterfactuals, causal sequences and explanation. In *The Psychology of Counterfactual Thinking*. 2005.

[Hilton, 1990] Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65–81, 1990.

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

[Kahneman and Tversky, 1982] Daniel Kahneman and Amos Tversky. The simulation heuristic. In P. Slovic D. Kahneman and A. Tversky, editors, *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press, 1982.

[Kelley, 1967] Harold H Kelley. Attribution theory in social psychology. In *Nebraska symposium on motivation*, pages 192–238. Uni. Nebraska Press, 1967.

[Lipton, 1990] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990.

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

[Lombrozo, 2012] Tania Lombrozo. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276, 2012.

[Malle, 2004] Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press, 2004.

[Malle, 2011] Bertram F Malle. Time to give up the dogmas of attribution: An alternative theory of behavior explanation. *Advances in Experimental Social Psychology*, 44(1):297–311, 2011.

[McCloy and Byrne, 2000] Rachel McCloy and Ruth MJ Byrne. Counterfactual thinking about controllable events. *Memory & Cognition*, 28(6):1071–1078, 2000.

[Miller and Gunasegaram, 1990] Dale T Miller and Saku Gunasegaram. Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of personality and social psychology*, 59(6):1111, 1990.

[Miller, 2017] Tim Miller. Explainable AI: Insights from the social sciences. *ArXiv e-prints*, 1706.07269, 2017. `https://arxiv.org/abs/1706.07269`.

[Nott, 2017] George Nott. 'Explainable Artificial Intelligence': Cracking open the black box of AI. *Computer World*, 4 2017. `https://www.computerworld.com.au/article/617359/`.

[Rao and Georgeff, 1995] Anand S Rao and Michael P Georgeff. Bdi agents: From theory to practice. In *ICMAS*, volume 95, pages 312–319, 1995.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the Int. Conf.x on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[Slugoski *et al.*, 1993] Ben R Slugoski, Mansur Lalljee, Roger Lamb, and Gerald P Ginsburg. Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology*, 23(3):219–238, 1993.

[Swartout and Moore, 1993] William R Swartout and Johanna D Moore. Explanation in second generation expert systems. In *Second generation expert systems*, pages 543–585. Springer, 1993.

[Thagard, 1989] Paul Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12(03):435–467, 1989.

## A List of Papers Surveyed

Taken from the 'Related Work' list posted on the website for the IJCAI 2017 Explainable AI workshop[2] as of 16 May 2017.


1. Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S. (2017). Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. To appear in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press.
2. Cheng, H., et al. (2014) SRI-Sarnoff Aurora at TRECVid 2014: Multimedia event detection and recounting.
3. Doshi-Velez, F., & Kim, B. (2017). A roadmap for a rigorous science of interpretability. (arXiv:1702.08608)
4. Elhoseiny, M., Liu, J., Cheng, H., Sawhney, H., & Elgammal, A. (2015). Zero-shot event detection by multimodal distributional semantic embedding of videos. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (pp. 3478-3486). Phoenix, AZ: AAAI Press.
5. Hendricks, L.A, Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. (arXiv:1603.08507v1)
6. Kofod-Petersen, A., Cassens, J., & Aamodt, A. (2008). Explanatory capabilities in the CREEK knowledge-intensive case-based reasoner. Frontiers in Artificial Intelligence and Applications, 173, 28-35.
7. Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. Proceedings of the Twentieth International Conference on Intelligent User Interfaces (pp. 126-137). Atlanta, GA: ACM Press.
8. Lake, B.H., Salakhutdinov, R., & Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. Science, 350, 1332-1338.
9. Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable agency for intelligent autonomous systems. In Proceedings of the Twenty-Ninth Annual Conference on Innovative Applications of Artificial Intelligence. San Francisco: AAAI Press.
10. Lécué, F. (2012). Diagnosing changes in an ontology stream: A DL reasoning approach. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. Toronto, Ontario, Canada: AAAI Press.
11. Letham, B., Rudin. C., McCormick, T., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics, 9(3), 1350-137.
12. Lombrozo, T. (2012). Explanation and abductive inference. Oxford Handbook of Thinking And Reasoning (pp. 260-276).
13. Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. MIS Quarterly, 38(1), 73-99.
14. Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Human Centered Machine Learning: Papers from the CHI Workshop. (arXiv:1602.04938v1)
15. Rosenthal, S., Selvaraj, S. P., & Veloso, M. (2016). Verbalization: Narration of autonomous mobile robot experience. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, NY: AAAI Press.
16. Sheh, R.K. (2017). "Why did you do that?" Explainable intelligent robots. In K. Talamadupula, S. Sohrabi, L. Michael, & B. Srivastava (Eds.) Human-Aware Artificial Intelligence: Papers from the AAAI Workshop (Technical Report WS-17-11). San Francisco, CA: AAAI Press.
17. Si, Z. and Zhu, S. (2013). Learning AND-OR templates for object recognition and detection. IEEE Transactions On Pattern Analysis and Machine Intelligence, 35(9), 2189-2205.
18. Shwartz-Ziv, R. & Tishby, N. (2017). Opening the black box of deep neural networks via information. (arXiv:1703.00810 [cs.LG])
19. Sormo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning: Perspectives and goals. Artificial Intelligence Review, 24(2), 109-143.
20. Swartout, W., Paris, C., & Moore, J. (1991). Explanations in knowledge systems: Design for explainable expert systems. IEEE Expert, 6(3), 58-64.
21. van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. Proceedings of the Nineteenth National Conference on Artificial Intelligence (pp. 900-907). San Jose, CA: AAAI Press.
22. Zahavy, T., Zrihem, N.B., & Mannor, S. (2017). Graying the black box: Understanding DQNs. (arXiv:1602.02658 [cs.LG])
23. Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H.H., & Kambhampati, S. (2017). Plan explicability and predictability for robot task planning. To appear in Proceedings of the International Conference on Robotics and Automation. Singapore: IEEE Press.


[2]See `http://home.earthlink.net/~dwaha/research/meetings/ijcai17-xai/`.

## B Detailed Results

| Paper | On topic | Data Driven | Validation | Comments |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | |
| 2 | 0 | 0 | 0 | |
| 3 | 1 | 1 | N/A | A position paper, so Validation not applicable. |
| 4 | 0 | 0 | 0 | |
| 5 | 1 | 0 | 1 | |
| 6 | 1 | 1 | 0 | |
| 7 | 1 | 0 | 1 | |
| 8 | 0 | 2 | 1 | Off topic, but is mature work |
| 9 | 1 | 0 | N/A | |
| 10 | 0 | 0 | 0 | |
| 11 | 1 | ? | 0 | Could not locate reference Jennings et al. (1982) |
| 12 | N/A | N/A | N/A | Survey paper on explanation in the social sciences |
| 13 | 1 | 0 | 0 | |
| 14 | 1 | 0 | 1 | |
| 15 | 1 | 0 | 0 | |
| 16 | 1 | 0 | 0 | |
| 17 | 0 | 0 | 0 | |
| 18 | 1 | 0 | 0 | |
| 19 | 1 | 2 | N/A | Survey paper, so Validation not applicable |
| 20 | 1 | 0 | 0 | |
| 21 | 1 | 0 | 1 | |
| 22 | 1 | 0 | 0 | |
| 23 | 1 | 1 | 0 | |