Title

Explanation in Human Thinking

Permalink

https://escholarship.org/uc/item/9k6291nk

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

Authors

Cassens, Jörg
Habenicht, Lorenz
Blohm, Julian
et al.

Publication Date

2021

Explanation in Human Thinking

Jörg Cassens (joerg.cassens@uni-hildesheim.de)
Lorenz Habenicht (haben002@uni-hildesheim.de)
Julian Blohm (blohmj@uni-hildesheim.de)
University of Hildesheim, Hildesheim, Germany

Rebekah Wegener
(rebekah.wegener@sbg.ac.at)
University of Salzburg, Salzburg, Austria

Joanna Korman (jkorman@bentley.edu)
Sangeet Khemlani (sunny.khemlani@nrl.navy.mil)
US Naval Research Laboratory
Washington D.C.,
United States

Giorgio Gronchi
(giorgio.gronchi@unifi.it)
University of Florence,
Florence, Italy

Ruth M.J. Byrne (rmbyrne@tcd.ie)
Trinity College Dublin,
Dublin, Ireland

Greta Warren (greta.warren@ucdconnect.ie)
Molly S. Quinn (molly.quinn@ucdconnect.ie)[*]
Mark T. Keane (mark.keane@ucd.ie)[*]
School of Computer Science & VistaMilk SFI Centre,
University College Dublin, Dublin, Ireland

Keywords: explanation; dialogic; explanatory reasoning; incompleteness; illusionism; counterfactuals; case-based explanations

Explaining as Dialogic Process

Jörg Cassens, Rebekah Wegener, Lorenz Habenicht, and Julian Blohm discuss the dialogic form of explanations. Explanations are a long established research topic in a wide variety of disciplines, ranging from philosophy (van Fraassen, 1980; Achinstein, 1983) over the cognitive sciences and psychology (Lalljee et al., 1983; Keil and Wilson, 2000; Lombrozo, 2006) to computer science in general and artificial intelligence in particular (Schank, 1986; Leake, 1992; Leake (1995); Sørmo et al., 2005). However, while there is compelling research supporting the value, structure and function of explanation, as Edwards et al. (2019) argue, "accounts of explanation typically define explanation (the product) rather than explaining (the process)". By contrast, we aim at an understanding of explanation as a functional variety of language behaviour that treats explanations as being

- **Contextualised**, which itself is comprised of a) Context Awareness (knowing the situation the system is in) and b) Context Sensitivity (acting according to such situation),
- **Construed by user interest**,
- **Multimodal**, and
- **Dialogic**.

In this talk, we will focus on the latter two aspects, the dialogic form of explanations and its representation in different modalities (and codalities). We will report on our recent empirical work where we have been looking at explanatory situations, firstly the differences between human to human explanations and machine to human explanations (using of-the-shelf speech dialogue systems) and secondly multimodal human to human explanations.

How People Judge the Completeness of an Explanation

Joanna Korman and Sangeet Khemlani investigate how people judge the completeness of explanations. All explanations are incomplete descriptions, but reasoners appear willing to assess some explanations as more complete than others. To explain this behavior, we propose a novel theory of the detection of explanatory incompleteness. The account assumes that reasoners represent explanations as mental models – discrete, iconic representations of possibilities – of causal scenarios. A complete explanation refers to a single integrated model, whereas an incomplete explanation refers to multiple models. The theory predicts that if there exists an unspecified causal relation – a gap – anywhere within a causal description, then reasoners will maintain multiple explanatory models to handle the gap (Korman & Khemlani, 2020). Reasoners should treat such explanations as less complete than those without such a gap. Four experiments provided participants with causal descriptions, some of which yield one explanatory model, e.g., A causes B and B causes C, and some of which demand multiple models, e.g., A causes X and B causes C. Participants across the studies preferred one-model descriptions to multiple-model ones on tasks that implicitly and explicitly required them to assess explanatory completeness. The studies corroborate the theory, and they are the first to reveal the processes that underlie the assessment of explanatory completeness. We conclude by reviewing the theory in light of extant accounts of causal reasoning.

[*] Co-organizers.

The Potential of Illusionism for Investigating Explanation in Human Thinking

Giorgio Gronchi describes the potential of illusionism for investigating explanation in human thinking. In the last decade, a research program that uses magical effects to study psychological phenomena has emerged (Kuhn, Olson & Raz, 2016). This approach has mainly focused on perception and attention but magicians' techniques and experiences can also give clues to understand human thinking. This talk will focus on illusionism as a tool to investigate how people explain impossible phenomena. Also, magicians' literature may offer new insights. For example, magicians employ systematically the explaining away properties and magical effects may be employed to investigate inverted forks causal relationship. Indeed, the impossible effects produced by magicians have an actual cause (the trick) but in order to make the trick less understandable, magicians offer manifestly a false explanation (e.g., the use of a pinch of magical powder). This is crucial for the success of the magical performance. Illusionism literature implies that this procedure works even if there is not an actual causal relationship between the false cause (a wave of magical wand to find a card in a deck) and the apparently impossible effect, especially according to modern common-sense. Empirical observations in this respect will be discussed also in light of the different type of cover stories (the more or less explicit explanation suggested by the magician during the act) employed by magicians depending if the impossible effect Is physical-based (such as somebody cut in half and put together again) or mind-based (e.g., predicting a choice with the power of mind).

Counterfactual Explanations in Explainable AI

Greta Warren, Molly Quinn, Ruth Byrne, and Mark Keane discuss how counterfactuals can be used to provide better case-based explanations for AI-systems. We report the results of experiments to test the goodness of explanations for the performance of an artificial intelligence (AI) system. AI systems designed to support human decision-making are becoming increasingly prevalent and yet their decisions are complex for human users to understand. To increase trust and acceptance by a human user of an AI system needs to adequately explain its decisions. Often human users of such systems engage in counterfactual and contrastive thinking – 'why this decision and not that one?' People tend to create counterfactuals that focus on certain "fault-lines" such as exceptions to norms. Our experiments examine the sorts of counterfactual explanations that people find useful for AI systems that perform a variety of prediction /classification tasks (e.g., predicting blood alcohol levels, classifying written numbers). The AI system relies on case-based explanations and can generate nearest-neighbor cases as explanations for a classification/prediction, as well as counterfactual cases, with a different outcome (i.e., nearest-unlike-neighbor cases). We test the effects of these different sorts of case-based explanations on judgments of the adequacy, accuracy and trust in the AI system. We discuss the implications of the results for understanding cognitive processes in human explanation and argumentation, as well as their implications for understanding counterfactual thinking.

References

Achinstein, P. (1983). *The nature of explanation.* Oxford University Press, Oxford.

Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019) Explanation recruits comparison in a category-learning task. *Cognition, 185*, 21–38.

Halliday, M. A. K. (1978). *Language as a Social Semiotic: The Social Interpretation of Language and Meaning.* University Park Press.

Keil, F. C. & Wilson, R. A. (2000). Explaining Explanation. In *Explanation and Cognition,* (pp. 1–18). Bradford Books.

Korman, J., & Khemlani, S. (2020). Explanatory completeness. *Acta Psychologica, 209,* 103139.

Kuhn, G., Olson, J. A., & Raz, A. (2016). The psychology of magic and the magic of psychology. *Frontiers in Psychology, 7,* 1358.

Lalljee, M., Watson, M., & White, P. (1983) Attribution theory: Social and functional extensions. In *The Organization of Explanations.* Blackwell, Oxford.

Leake, D. B. (1992). *Evaluating Explanations: A Content Theory.* Lawrence Erlbaum Associates, New York.

Leake, D. B. (1995). Goal-based explanation evaluation. In *Goal-Driven Learning,* (pp. 251–285). MIT Press, Cambridge.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*(10), 464–470.

Schank, R. C. (1986). *Explanation Patterns – Understanding Mechanically and Creatively.* Lawrence Erlbaum, New York.

Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning– perspectives and goals. A*rtificial Intelligence Review*, *24*(2), 109–143. ISSN 0269-2821.

van Fraassen, B. C. (1980). *The Scientific Image.* Clarendon Press, Oxford.