Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Doctoral Consortium Track

Predictive Modelling of Human Reasoning Using AGM Belief Revision

Clayton Baker
University of Cape Town, South Africa
bkrcla003@myuct.ac.za

Abstract

While many forms of belief change exist, the relationship between belief revision and human reasoning is of primary interest in this work. The theory of belief revision extends classical two-valued logic with an approach to resolve the conflict between a set of beliefs and newly learned information. The goal of this project is to test how humans revise conflicting beliefs. Experiments are proposed in which human subjects are required to resolve conflicting beliefs via relevance and confidence. In our analysis, the human responses will be evaluated against the predictions of two perspectives of propositional belief revision: formal and psychological.

1 Introduction

Johnson-Laird, renown philosopher of language and reasoning, holds the view that humans exploit all the knowledge available to them to arrive at an inference, rather than apply formal logical rules. In response, the task of making formal rules of inference flexible to exceptions and tolerant of new information has been studied under the term non-monotonic reasoning. A proponent [Oaksford and Chater, 2020] views reasoning as a social activity which, when formalised in classical true-false logic, is weakened by a lack of contextual knowledge and the limitation that background knowledge must be consistent. Both non-monotonic reasoning and belief change have been proposed as a better candidate for modelling human revision than classical logic. In the non-monotonic case, for example, suppose a reasoner has the knowledge that "Daisy is a Dodo", "Dodos are birds found around island nations in the Indian Ocean" and "birds fly'. Learning that "Dodos are flightless birds" presents evidence that causes the reasoner to withdraw the accepted knowledge that birds fly. On the other hand, belief change [Dubois *et al.*, 2020] is the study of how old beliefs are changed to accommodate new information. While many forms of belief change have been studied, the relationship between belief revision and human reasoning is of primary interest in this work. The theory of belief revision extends classical logic with an approach to resolving conflict between a set of beliefs and newly learned information. For example, suppose a reasoner holds the belief that "A mystery box is behind door A or door B". This belief can be represented by the conjunction $\alpha \vee \beta$, where α and β indicates that the mystery box is behind door A and door B, respectively. Learning that door A opens to an empty room, causes the reasoner to revise their beliefs with $\neg\alpha$, which means that the mystery box is not behind door A. To resolve the conflicting information, a reasoner may revise their beliefs by concluding that the mystery box is not behind door A, but behind door B, i.e. $\neg\alpha \wedge \beta$. The broad objective of this project is to understand more about how humans revise their beliefs when presented with conflicting information. In previous work [Baker and Meyer, 2021], an output of Master's research, we surveyed English translations of the 8 Alchourrón, Gärdenfors and Makinson (AGM) [Alchourrón *et al.*, 1985] postulates of belief revision with human subjects. Half of the postulates were found agreeable with the subjects. A limitation of this work is that only one translation per postulate was used and that a comparison of the data for each postulate was not telling. Since humans are known to construct visual representations of their environment, it would be necessary to test each postulate repeatedly using varied material. Another limitation is that although the AGM postulates are a sound framework to examine the relationship between human reasoning and belief revision, the postulates are not considered complete. In recent literature, the AGM theory has been developed through the notions of relevance and independence, in which only the part of a belief set that is affected by the new information should be revised. Syntax splitting by Parikh [1999], considers this key idea concerning revising belief sets. In later work, Aravanis et al. [Aravanis *et al.*, 2019] specified epistemic entrenchment models [Gärdenfors and Makinson, 1988] and AGM-style partial meet models for Parikh's relevance-sensitive axiom. Identified as a desirable property of knowledge representation and reasoning, syntax splitting has also been extended to the settings of iterated revision [Kern-Isberner and Brewka, 2017], non-monotonic reasoning [Kern-Isberner *et al.*, 2020] and contraction [Haldimann *et al.*, 2020], amongst others. While the revision theory has been extensively developed for enhancing logical inference, it is less clear whether humans find revision natural and useful. When incorporating new information into an existing belief set, a revision operation ranks the set of interpretations using total pre-orders [Delgrande *et al.*, 2018]. The interpretations (beliefs) that are

closest to the old set of beliefs are included in the revised set. We ask whether humans apply a similar ranking process to their beliefs when learning new information. In addition, a general population of human subjects may agree on a limited set of beliefs, especially when there is conflict. We propose to investigate whether humans find consistent beliefs important, and how conflicting information is resolved. The kinds of beliefs that human subjects retain or discard when learning new information will be studied, alongside the context of the situation. Furthermore, the AGM account of belief revision is based on logic, while cognitive models do not have a logical core. Instead, cognitive models use a representation of someone's intuition, knowledge, and interactions with the world. Johnson's [2010] is such a cognitive model that uses propositions to denote beliefs. The model makes deductions based on propositions, e.g "the car is parked outside", using *if*, *and*, *or* and *not*. We will draw a comparison between these two perspectives concerning their predictive power on human responses to resolving conflicting beliefs. A core requirement of this project is collecting data from human subjects. We propose to do this via two experiments: one to test how subjects identify information that is relevant in a given context and the other to test how subjects esteem beliefs after learning new information. The data from both surveys will be combined to produce a full representation of the subjects' beliefs. The variables of interest in the experiment are information relevance and information confidence. For each conversation, we will build two corresponding representations. The responses will first be analysed using standard statistical tests. Next, we will determine whether MMT can predict individual responses. Conveniently, MMT is a built-in model an open-source software platform called Cognitive Computation for Behavioural Reasoning Analysis (CCOBRA) [1]. The inferential power of MMT, via CCOBRA, will be compared to the logical inference of the 8 AGM postulates and desirable postulates for information relevance [Peppas and Williams, 2016]. In future, we will consider the impact of the broader work in the social science literature. A long-term goal of this work is to add significant features of revision to the existing MMT model on CCOBRA.

Ethical Statement

The ethical statement for this project can be accessed here: https://tinyurl.com/2p946xty.

Acknowledgments

Thanks to my supervisor, Prof. Thomas Meyer, for his mentorship and supervision of this project. Guidance on the cognitive science aspect was provided by Prof. Marco Ragni.

References

[Alchourrón *et al.*, 1985] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[Aravanis *et al.*, 2019] Theofanis Aravanis, Pavlos Peppas, and Mary-Anne Williams. Full characterization of parikh's relevance-sensitive axiom for belief revision. *Journal of Artificial Intelligence Research*, 66:765–792, 2019.

[Baker and Meyer, 2021] C.K. Baker and T. Meyer. Belief change in human reasoning: An empirical investigation on mturk. In *Proceedings of the Second Southern African Conference for Artificial Intelligence Research (SACAIR 2021)*, pages 520–536, 2021.

[Delgrande *et al.*, 2018] James P Delgrande, Pavlos Peppas, and Stefan Woltran. General belief revision. *Journal of the ACM (JACM)*, 65(5):1–34, 2018.

[Dubois *et al.*, 2020] Didier Dubois, Patricia Everaere, Sébastien Konieczny, and Odile Papini. Main issues in belief revision, belief merging and information fusion. *A Guided Tour of Artificial Intelligence Research: Volume I: Knowledge Representation, Reasoning and Learning*, pages 441–485, 2020.

[Gärdenfors and Makinson, 1988] Peter Gärdenfors and David Makinson. Revisions of knowledge systems using epistemic entrenchment. In *Proceedings of the 2nd conference on Theoretical aspects of reasoning about knowledge*, pages 83–95, 1988.

[Haldimann *et al.*, 2020] Jonas Philipp Haldimann, Gabriele Kern-Isberner, and Christoph Beierle. Syntax splitting for iterated contractions. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 465–475, 2020.

[Johnson-Laird, 2010] Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

[Kern-Isberner and Brewka, 2017] Gabriele Kern-Isberner and Gerhard Brewka. Strong syntax splitting for iterated belief revision. In *IJCAI*, pages 1131–1137, 2017.

[Kern-Isberner *et al.*, 2020] Gabriele Kern-Isberner, Christoph Beierle, and Gerhard Brewka. Syntax splitting= relevance+ independence: New postulates for nonmonotonic reasoning from conditional belief bases. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 560–571, 2020.

[Oaksford and Chater, 2020] Mike Oaksford and Nick Chater. New paradigms in the psychology of reasoning. *Annual review of psychology*, 71:305–330, 2020.

[Parikh, 1999] Rohit Parikh. Beliefs, belief revision, and splitting languages. *Logic, language and computation*, 2(96):266–268, 1999.

[Peppas and Williams, 2016] Pavlos Peppas and Mary-Anne Williams. Kinetic consistency and relevance in belief revision. In *Logics in Artificial Intelligence: 15th European Conference, JELIA 2016, Larnaca, Cyprus, November 9-11, 2016, Proceedings 15*, pages 401–414. Springer, 2016.

[1] https://orca.informatik.uni-freiburg.de/ccobra