







Cognitive Defeasible Reasoning: the Extent to Which Forms of Defeasible Reasoning Correspond with Human Reasoning

Clayton Kevin Baker^(✉) , Claire Denny , Paul Freund ,
and Thomas Meyer 

University of Cape Town and CAIR, Cape Town, South Africa
{bkrcla003,dnncla004,frnpau013}@myuct.ac.za, tmeyer@cs.uct.ac.za

Abstract. Classical logic forms the basis of knowledge representation and reasoning in AI. In the real world, however, classical logic alone is insufficient to describe the reasoning behaviour of human beings. It lacks the flexibility so characteristically required of reasoning under uncertainty, reasoning under incomplete information and reasoning with new information, as humans must. In response, non-classical extensions to propositional logic have been formulated, to provide non-monotonicity. It has been shown in previous studies that human reasoning exhibits non-monotonicity. This work is the product of merging three independent studies, each one focusing on a different formalism for non-monotonic reasoning: KLM defeasible reasoning, AGM belief revision and KM belief update. We investigate, for each of the postulates propounded to characterise these logic forms, the extent to which they have correspondence with human reasoners. We do this via three respective experiments and present each of the postulates in concrete and abstract form. We discuss related work, our experiment design, testing and evaluation, and report on the results from our experiments. We find evidence to believe that 1 out of 5 KLM defeasible reasoning postulates, 3 out of 8 AGM belief revision postulates and 4 out of 8 KM belief update postulates conform in both the concrete and abstract case. For each experiment, we performed an additional investigation. In the experiments of KLM defeasible reasoning and AGM belief revision, we analyse the explanations given by participants to determine whether the postulates have a normative or descriptive relationship with human reasoning. We find evidence that suggests, overall, KLM defeasible reasoning has a normative relationship with human reasoning while AGM belief revision has a descriptive relationship with human reasoning. In the experiment of KM belief update, we discuss counter-examples to the KM postulates.

Keywords: Non-monotonic reasoning · Defeasible reasoning · Belief revision · Belief update · Survey · Google forms · Mechanical turk

Supported by Centre for Artificial Intelligence Research (CAIR).

© Springer Nature Switzerland AG 2020
A. Gerber (Ed.): SACAIR 2020, CCIS 1342, pp. 199–219, 2020.
https://doi.org/10.1007/978-3-030-66151-9_13

1 Introduction

It is well-documented that human reasoning exhibits flexibility considered key to intelligence [21], yet fails to conform to the prescriptions of classical or propositional logic [25] in the Artificial Intelligence (AI) community. The AI community, therefore, seeks to incorporate such flexibility in their work [21]. Non-classical or non-monotonic logic is flexible by nature. Whereas classical reasoning is enough to describe systems with a calculated output in an efficient way, how humans reason is non-classical because humans are known to reason in different ways [25]. The problem is that non-monotonic reasoning schemes have been developed for and tested on computers, but not on humans. There is a need to investigate whether there exists a correspondence between non-monotonic reasoning and human reasoning and, if so, to what extent it exists. This problem is important because we can gain insight into how humans reason and incorporate this into building improved non-monotonic AI systems. An issue which needs to be considered is that humans are diverse subjects: some reason normatively while others reason descriptively. In the case of normative reasoning, a reasoner would conclude that a certain condition *should be* the case or that the condition is usually the case. In the case of descriptive reasoning, a reasoner would make a bold claim that a certain condition *is* exactly true or exactly false. We emphasise that a thorough investigation needs to be done to determine the extent of the correspondence between non-monotonic reasoning and how humans reason.

We propose this work as a contribution towards solving this problem. While acknowledging that this work falls within a broader research paradigm towards this goal [21, 24, 25], what differentiates this work is that it is, to our knowledge, the first work with an explicit view towards testing each of these particular formal non-monotonic frameworks: KLM defeasible reasoning [13], AGM belief revision [1], and KM belief update [11]. We report on these frameworks in a paper due to the close theoretical links between the frameworks' domains. Postulates for defeasible reasoning and belief revision may be translated from the one context to the other [5]. Using such translations, KLM defeasible reasoning [13] can be shown to be the formal counterpart of AGM belief revision [8]. This does not hold for KM belief update [11]. Belief update is commonly considered a necessarily distinct variant of belief revision for describing peoples' beliefs in certain domains [11].

In Sect. 2, we describe related work and the formalisms of non-monotonic reasoning under investigation in our study. We end this section with our problem statement. In Sect. 3, we describe the design and implementation of three distinct surveys, one for each formalism of non-monotonic reasoning in our study. Each survey seeks to determine the extent of correspondence between the postulates of that formalism and human reasoning. In Sect. 4, we describe the methods used to analyse our survey results. We present our results, discussion and conclusions in Sect. 5. Lastly, we propose the track for future work in Sect. 6.

2 Background

Humans are known to reason differently about situations in everyday life and this reasoning behaviour can be compared to the paradigm of non-monotonic reasoning in AI. Non-monotonic reasoning is the study of those ways of inferring additional information from given information that does not satisfy the monotonicity property, which is satisfied by all methods based on classical logic [13]. Said otherwise, non-monotonic logic fails the principle that whenever x follows from a set A of propositions then it also follows from every set B with $B \subseteq A$ [18]. With non-monotonic reasoning, a conclusion drawn about a particular situation does not always hold i.e. in light of newly gained, valid information, previously valid conclusions have to change. This type of reasoning is described in the context of AI [23]. We consider three forms of non-monotonic reasoning, namely defeasible reasoning, belief revision and belief update. The latter two are both forms of *belief change* [11], wherein there exists a belief base and a belief set [6]. Explicit knowledge the agent has about the world resides in the belief base, whereas both the explicit knowledge the agent has about the world and the inferences derived from it reside in the belief set.

2.1 Defeasible Reasoning

Defeasible reasoning occurs when the evidence available to the reasoner does not guarantee the truth of the conclusion being drawn [21]. A defeasible statement has two identifiable parts: an antecedent or premises and a consequence or conclusion [7]. With classical reasoning, we proceed from valid premises to a valid conclusion and this conclusion will never change. With defeasible reasoning, we can proceed from valid premises to a valid conclusion also. However, in light of new valid information, the previously valid conclusion is allowed to change. Either the conclusion will be supported by the new information or the conclusion will be defeated by the new information. This defeasible reasoning behaviour applies to many aspects of the everyday life of humans, where the information available to the reasoner is often incomplete or contains errors. As such, defeasible information often involves information that is considered typical, normal or plausible. We shall now illustrate this with an example. Consider the following statements: *employees pay tax* and *Alice is an employee*. From the statements given, can we conclude that *Alice pays tax*? In the classical case, we can only conclude that *Alice pays tax*. Using defeasible reasoning, we can also conclude that *Alice pays tax*. However, should we receive additional information about Alice that she is not a typical employee, we can change our conclusion to be *Alice does not pay tax*. In this case, we have to amend our premises to account for the defeasible information viz. *employees typically pay tax* and *Alice is not a typical employee*.

2.2 Belief Revision

In belief revision, conflicting information indicates flawed prior knowledge on the part of the agent, forcing the retraction of conclusions drawn from

it [11, 19]. Information is then taken into account by selecting the models of the new information closest to the models of the base, where a model of information μ is a state of the world in which μ is true [11]. An example of this reasoning pattern will now be described. Consider the same statements used above in the defeasible reasoning example. Using the reasoning pattern of belief revision, we can infer from our beliefs that Alice does pay tax. Suppose we now receive new information: *Alice does not pay tax*. This is inconsistent with our belief base, so a decision must be made regarding which beliefs to retract prior to adding the new information into our beliefs. We could revise our beliefs to be that *employees pay tax* and *Alice does not pay tax*. In [4], this decision is proposed to be influenced by whether we believe some statements more strongly than others. In [1], it is proposed to be influenced by closeness (the concept of minimal change), in that we aim to change as little about our existing knowledge as we can do without having conflicting beliefs.

2.3 Belief Update

In belief update, conflicting information is seen as reflecting the fact that the world has changed, without the agent being wrong about the past state of the world. To get an intuitive grasp of the distinction between belief update and revision, take the following example adapted from [11]. Let b be the proposition that the book is on the table, and m be the proposition that the magazine is on the table. Say that our belief set includes $(b \wedge \neg m) \vee (\neg b \wedge m)$, that is the book is on the table or the magazine is on the table, but not both. We send a student in to report on the state of the book. She comes back and tells us that the book is on the table, that is b . Under the AGM [1] postulates for belief revision proposed in [1], we would be warranted in concluding that $b \wedge \neg m$, that is, the book is on the table and the magazine is not. But consider if we had instead asked her to ensure that the book was on the table. After reporting, we again are faced with the new knowledge that b . This time adding the new knowledge corresponds to the case of belief update. And here it seems presumptuous to conclude that the magazine is not on the table [11]. Either the book was already on the table and the magazine was not, in which case the student would have done nothing and left, or the magazine was on the table and the book not, in which case the student presumably would have simply put the book on the table and left the magazine similarly so. As these examples are formally identical, there is a need for different formalisms to accommodate both cases.

2.4 Problem Statement

We propose a first study to address the gap between the postulates of KLM [13] defeasible reasoning, AGM [1] belief revision and KM [11] belief update, and human reasoning.

Research Question: To what extent do the postulates of defeasible reasoning, belief revision and belief update correspond with human reasoning?

We have investigated three approaches to non-monotonic reasoning: the KLM [13] defeasible reasoning approach, the AGM [1] belief revision approach and the KM [11] belief update approach. In additional investigations, the reasoning style of participants, normative or descriptive, was identified in the cases of defeasible reasoning and belief revision. For belief update, the additional investigation was to find counter-examples to the KM [11] postulates.

3 Implementation

In this section, we describe the design and implementation of three surveys: one each for defeasible reasoning, belief revision, and belief update. We also describe our implementation strategy and expected challenges. Finally, we document our testing and evaluation strategy. The major reason for our choice of the survey as a testing instrument was its ease of integration with Mechanical Turk, which was the channel we had chosen for sourcing our participants. Moreover, the web-based survey is a common tool used in sociological research, such that “it might be considered an essential part of the sociological toolkit” [32]. Future work may look towards testing our research questions in a non-survey environment.

3.1 Survey Designs

Each of the three surveys focused on testing a particular formalism of non-monotonic reasoning: survey 1 tested defeasible reasoning, survey 2 tested belief revision and survey 3 tested belief update. 30 responses were wanted per survey. Participants were asked whether they accepted the conclusions proposed by the postulates of the formalism of non-monotonic reasoning in question, and were asked to give an explanation for their answer. The postulates were presented in concrete and abstract form. The concrete part of the survey consisted of the translations of all the postulates into English sentences. The abstract part consisted of the translations of all the postulates into variables, denoted by capital letters from the English alphabet. The logical behaviour of the postulates was maintained in the translations of these postulates from propositional logic to their concrete and abstract forms, respectively. For a particular postulate in the concrete case, the premises and conclusion were substituted with sentences from the English language. In the abstract case, the premises and conclusion were substituted by variables using letters from the English alphabet. Together, the premises and conclusion for each postulate created a story for the participant to read. The stories used in the concrete part of the survey were designed to mimic an environment in which a general reasoner might find himself. For example, some of the stories related to reasoning about students and homework, whilst others related to reasoning about the weather. The stories used in the abstract part were less verbose as no context was given to indicate the meaning of the variables used. An example of a concrete, story-style or real-world question would be: *If Cathy has a cake to bake, will she use an oven?* An example of an abstract question would be: given the following, *If A then B*, and *If C then A*, can we

say that *If C then B*? The survey questions can be navigated to by means of Appendix A for reference.

Survey 1. This survey tested participants' ability to reason defeasibly. Participants were asked whether they accepted the conclusions proposed by the KLM [13] postulates of defeasible reasoning and were required to provide explanations for their reasoning. We refer to the KLM [13] postulates of *Left Logical Equivalence* (LLE), *Right Weakening* (RW), *And*, *Or* and *Cautious Monotonicity* (CM), included in Appendix A for reference. The KLM [13] postulates were presented as textual stories containing a set of information, or premises, and a proposed conclusion. For each postulate, the stories were included in concrete and abstract form. The concrete form of the postulates was kept separately from the abstract form. The concrete part of the survey was presented to participants first. The abstract part was presented next. For both the concrete and abstract parts, the order of the postulates was randomised. Crucial to this study, the explanations given by participants were used to identify whether they reasoned normatively or descriptively. This survey also tested participants' ability to reason defeasibly in a broader sense. In particular, additional defeasible reasoning postulates were presented to participants. This was done by presenting each postulate in concrete and abstract form and asking participants to reason as before. In the concrete case only, participants' ability to reason under two distinctive subcategories of defeasible reasoning, prototypical reasoning and presumptive reasoning, was tested. Prototypical reasoning [17] suggests each reasoning scenario assumes a model with certain typical features, whereas presumptive reasoning [31] suggests that an argument may have multiple possible consequences. As an avenue for future work, the ability for participants to reason prototypically and presumptively could be tested in greater detail with scope to include testing participants' ability to reason in the abstract case.

Survey 2. The questions in this survey were developed to test whether postulates of a specific formalisation of the process of belief revision feature in cognitive reasoning. The formalisation used is that of the eight-postulate approach as proposed by Alchourrón, Gärdenfors and Makinson (AGM) [1]. We refer to the eight-postulate approach as the AGM [1] postulates of *Closure*, *Success*, *Inclusion*, *Vacuity*, *Consistency*, *Extensionality*, *Super-expansion* and *Sub-expansion*, included in Appendix A for reference. Two types of questions were developed: concrete and abstract. This involved designing scenarios in which to ground the concrete questions. Five such scenarios were designed. Abstract questions were developed directly based on the formal postulates. The abstract questions were included to test the postulates without having the agent's knowledge of the world hindering their answers and to have questions which are less semantically loaded [16] than real-world concrete questions. The benefit of abstract examples is further discussed by Pelletier and Elio [21]. The concrete questions started as abstract representations explicitly requiring the application of one or some of the formal postulates to obtain the desired answer. These representations were then

elaborated in the context of a scenario. The scenarios designed are: linguists, smoking, wildlife, bag of stationery and, acrobats. The scenarios designed are inspired by the literature and the researcher’s knowledge of the world.

Survey 3. The questions in this survey were developed to test the KM approach [11] to belief update. The KM [11] postulates we used are included as postulates $U1$, $U2$, $U3$, $U4$, $U5$, $U6$, $U7$ and $U8$ in Appendix A. These postulates mirrored the eight-postulate approach for belief revision, with the core difference between the postulates for revision and the postulates for update being the type of knowledge referred to: static knowledge for revision and dynamic knowledge for update. The questions in this survey were broken into three sets. The first consisted of abstract questions, in which the KM [11] postulates were presented and participants were asked to rate their agreement with the postulates on a linear or Likert scale with extremal points “strongly agree” and “strongly disagree”. The postulates were presented using non-technical language. The second set of questions were concrete questions that were meant to be confirming instances of each of the eight KM postulates, where participants were asked to answer either *Yes* or *No*, and motivate their answer. The third set followed the same format as the second but was meant to present counter-examples to the postulates, with the counter-examples largely sourced from the literature. The first counter-example was based on the observation that updating p by $p \vee q$ does not affect the KM approach [9], which seems counter-intuitive. The second was based on the observation that updating by an inclusive disjunction leads to the exclusive disjunction being believed in the right conditions (a modification of the checkerboard example in [9]), which again seems counter-intuitive. The third was based on the observation that sometimes belief revision semantics seem appropriate in cases corresponding to the way that belief update is commonly, and has been here, presented in [15]. The final is an example testing a counter-intuitive result of treating equivalent sentences as leading to equivalent updates.

3.2 Mechanical Turk

Mechanical Turk (MTurk) is a service provided by Amazon that serves as an interface between service *Requesters* and a network of *Workers*. It addresses three problems [10]. It is used by software developers to incorporate human intelligence into software applications. It is used by business people to access a large network of human intelligence to complete tasks such as conducting market research. It is used by people looking to earn money to find work that can be done anywhere and at any time, using skills they already have. We used MTurk for access to its network of people to complete our surveys, which were hosted on Google Forms. An advantage is that its network of Workers includes people from a large range of ages, education levels and places [26]. Such places include the United States of America, India, Pakistan, the United Kingdom and the Philippines [26].

Although we did not set out to target a specific population of reasoners, MTurk offered a choice of up to 3 different qualifications that our Workers must

satisfy. For the defeasible reasoning survey, Workers were required to be Master Workers, a qualification assigned by MTurk to top Workers who consistently submit high-quality results. Workers were required to have a HIT Approval Rate (%) for all Requesters' HITs ≥ 97 , and have more than 0 HITs approved. For the belief revision survey, two MTurk qualifications and one internal qualification was used to recruit participants. Workers were required to have a HIT Approval Rate (%) for all Requesters' HITs > 98 , and have more than 5000 HITs approved. The required number of HITs approved was varied, between 1000 and 5000, to allow for a diverse sample of respondents. We created one internal qualification to ensure that the 30 respondents were unique across all of the published batches of the survey. This qualification was called *Completed my survey already* and assigned to Workers which have submitted a response in a previous batch, including the batch of the trial HIT. For the belief update survey, a single qualification was used: only Master Workers were allowed to participate in the survey.

3.3 Google Forms

Google Forms is an application which allows users to create and disseminate free online surveys. Research performed in 2018 revealed that the recent surge of low-quality qualitative data from MTurk is primarily due to international Turkers (workers on MTurk) [30] using Virtual Private Networks or Virtual Private Servers to waive qualifications required to complete surveys [12]. This motivated including a checkpoint within the surveys themselves, considering the surveys were answered online. The checkpoint comprised custom *captchas* and an attention check, designed to be an indicator of the respondent's suitability to take the survey. In this context, suitability comprises four requirements: (*i*) the response is not generated by a bot, (*ii*) the respondent is not using a script, (*iii*) the respondent can understand English, (*iv*) the respondent reads questions in full. Requirements *i* and *ii* address that the respondent must be a human. Requirement *iii* addresses that the survey questions are in English and require English answers. This limits the survey's population of potential respondents, as the respondent's English proficiency may affect their performance on the HIT e.g. in their interpretation of double negatives. A *Human Intelligence Task* (HIT) is any activity that can be performed on a computer by a human actor e.g. writing an essay. MTurk offered varying ages, backgrounds and other such contextual factors, resulting in it also presenting the challenge of verifying English proficiency levels, as understanding English is a broad classification.

We sought to clarify whether their understanding of a question posed in English was sufficient to answer correctly a trick question. We considered two options: to create a separate, qualifying HIT or include the qualifier as part of the survey. We chose the latter. Finally, requirement *iv* addresses that the respondent must be paying attention and reading all the information presented as prior knowledge before answering a question. If the respondent failed to meet requirement *iv*, we would lack cause to believe our assumption of what comprises their prior knowledge in later questions would not be violated.

3.4 Testing and Evaluation

Each of our surveys were evaluated by a group of both laypeople and experts for clarity. Each of our surveys were also published on MTurk as a trial HIT. The results of the trial HITs were used to gauge how Turkers might respond to the final survey.

Feedback from Groups of Laypeople and Experts. We asked a variety of experts and non-experts to evaluate our survey for coherence, clarity and other desirable characteristics of questions, more examples of which can be found in [14]. One of the authors evaluated each of the three surveys. We also approached an expert in psychology and an expert in philosophy, at the University of Cape Town, however they were not available to evaluate our surveys. The remaining experts who evaluated the survey questions included one Masters student in Computer Science, as well as two Computer Science Honours students also conducting studies on non-monotonic reasoning forms. One of the surveys was evaluated by an international doctoral student in language and African studies. Based on the suggestions from experts and laypeople, a variety of changes were made to the surveys.

Trial HITs. A trial of the surveys was conducted, (*i*) to gain familiarity with the MTurk service and platform and (*ii*) to test the survey and its questions on a sample of Turkers. It involved three separate postings of the survey links as HITs on the site, each requiring five responses. The HIT was created with certain specifications accordingly. Workers were compensated R30 (above the South African hourly minimum wage) for completing the tasks, and the tasks included a time estimate, all of which were under an hour. We did not restrict workers by location, but required that they should have completed a certain number of HITs previously, and have a certain approval rating ($\geq 95\%$) for their tasks, as recommended by Amazon to improve response quality [29]. A Turker's approval rating refers to the percentage of their tasks that have been approved or accepted by the Requesters who published them. Based on the results from the trial survey, changes were made for the final experiments. The changes included increasing both the compensation and the estimated completion time.

3.5 Ethical, Professional and Legal Issues

Ethical issues are those which require a choice to be made between options based on whether they evaluate as ethical or unethical. Professional issues here refer to those which pertain to ethical standards and rules that the profession of Computer Science has for its members, particularly with respect to research. Ethical and professional issues thus overlap. Legal issues refer to those which involve the law. As this project involved experiments with people, ethical clearance was obtained from the University of Cape Town Faculty of Science Human Research Ethics Committee before proceeding with the experiments. The primary issue

in the experiments was the use of MTurk, in particular, whether Workers were being paid a fair wage for their work. Per [2], the following three steps were taken to mitigate these concerns. First, workers were paid more than the South African minimum wage for an hour’s work. Second, in the title of the task, the estimated amount of time needed for the task was clearly stated. Finally, there is a section in the survey which gives an overview of what the research concerns, placing the work in context. Workers were also required to give their informed consent to participate in the study. This was achieved by having a consent form at the start of the survey, whereby workers could either agree to participate in the research and then continue to the rest of the survey, or they could decline to participate and be thanked for their time. Contact details of the researchers were also provided. Before the data-handling, all survey responses were anonymised. We also did not collect names, cellphone numbers or email addresses from our participants. The only personal contact information we collected from each participant was their Amazon Turk *WorkerID*. To view our survey questions, raw collected data and the codebooks used for data analysis, [click here](#).

4 Methods of Analysis

Responses were rejected if the participant failed the checkpoint section in the survey. In our analysis, we reference applying a baseline of 50% to our results. The choice of 50% as a baseline was arbitrary, but it served as a tool to evaluate the meaning of our results. As a starting point for evaluation and a baseline for agreement, it was basic and could be improved upon in future work.

4.1 Survey 1

The defeasible reasoning survey had 30 responses, which were downloaded from Google Forms. One response was rejected due to the participant submitting twice. Coding of participant responses was performed using Microsoft Excel functions. The coding spreadsheet is included in our Github repository, referenced in Appendix A. For this survey, we assumed that the KLM [13] postulate of *Reflexivity*, the idea that a proposition x defeasibly entails itself, holds for all human reasoners and therefore it was not tested. Feedback from our supervisor indicated that a few survey questions were not appropriate models of the KLM [13] postulates they intended to test, as they used the word *some* in the conclusion. These were questions 6 and 7, referring to the KLM [13] postulates of *Right Weakening* and *And*, respectively. In the following, we state question 6 as was presented in the survey, as an example, to clarify. The given information was presented as a numbered list and the conclusion was phrased as a question. Question 6, testing *Right Weakening*, asked: given i) no police dogs are vicious and ii) highly trained dogs are typically police dogs, can you conclude that *some* highly trained dogs are vicious? We draw the reader’s attention to the fact that the word *some* is not part of the definition of the KLM [13] postulates. Thus, we have removed the responses to these questions in our analysis of the results.

Quantitative Data. In our collected data, participant agreement with the postulates, the *Yes* or *No* responses, were considered quantitative data. This agreement was measured using a hit rate. The hit rate (%) for each postulate was calculated as with the formula: $\frac{\text{number of Yes responses}}{\text{total number of responses}} \times 100$. Hit rates were measured for each postulate in the concrete and the abstract case. A postulate with a hit rate of $\geq 50\%$ in both the concrete and abstract case was said to have agreement with the participants in this survey. We plotted the concrete and abstract hit rate for each defeasible reasoning postulate as well as the concrete hit rates for prototypical reasoning and presumptive reasoning, in Fig. 1. Where no data was available, this was indicated by a blank in the figure.

Qualitative Data. In our collected data, the explanations given by participants were considered qualitative data. We have identified four main emerging themes for participant explanations. The theme *Support* refers to an explanation which contained only information given in the question. The theme *Speculative* refers to an explanation for which there is partial support from the given information, but also in which external information, not present in the question, is considered. *Technical* refers to an explanation which contains the phrase *typically, but not always*. The theme *Other* refers to explanations which did not fit into any of the above categories and often contained explanations which were vague or explanations quoted from an external source but contained nonsensical words. After identifying these four themes, we have pooled the explanations from the *Support* and *Technical* themes and qualified these together as normative. We have qualified explanations fitting the *Speculative* theme as descriptive.

4.2 Survey 2

The belief revision survey had 40 participants, as 10 responses were rejected; 30 responses were used. Analysis of the Questions section of the survey, for both the trial and final survey, comprised finding the modal answer and hit rate for each closed question and performing qualitative analysis on the open questions. The data was downloaded from Google Forms and Mechanical Turk.

Quantitative Data. The modal answer and hit rate (%) for closed questions were calculated by applying Microsoft Excel functions to the data. A hit indicates success. In this context, success is defined as both the respondent and the application of the belief revision postulates obtaining the same answer. The hit rate is thus calculated for each question as $\frac{\text{number of successes}}{\text{no. of responses}} \times 100$. The analysis of the results employs a baseline of a hit rate of 50% to indicate overall success.

Qualitative Data. The qualitative analysis was performed in NVivo, a qualitative data analysis software package, and made use of *Tesch's Eight Steps in the Coding Process* [3]. In this process, a combination of pre-determined and emerging codes were used. Codes on topics expected to be found were taken

from literature, based on the theory being empirically tested. These include the eight postulates of belief revision as proposed by Alchourrón, Gärdenfors and Makinson [1]: closure, success, inclusion, vacuity, consistency, extensionality, super-expansion, sub-expansion. Other pre-determined codes include: normative and descriptive. Emerging codes are those which were not anticipated at the beginning, or are both unusual and of interest. They are developed solely on the basis of the data collected from respondents by means of the survey. An example of an emerging code used in the trial of this study is *It is stated*. This code represents the respondent taking a passive approach to their response. Other examples would be *real-world influence* and *likelihood*.

Pre-determined codes *normative* and *descriptive* refer to the reasoning style identified in responses to open questions. A normative style involves making value judgements [20], commenting on whether something is the way it should be or not. This includes implied judgements through the use of emotive language. A descriptive style, in contrast, does not - it involves making an observation, commenting on how something is [20].

4.3 Survey 3

Quantitative Data. The belief update survey had 34 participants, of which 4 responses were rejected. For the quantitative data, two forms of analysis were chosen, corresponding to the two different forms of quantitative data (ordinal and binary) gathered. For the ordinal (Likert-type) data, the median is an appropriate measure of central tendency [28], and thus was chosen, and for the binary data, the hit rate as above was chosen. Relating this back to the research question, a postulate was seen as confirmed if it saw both a hit rate $\geq 50\%$ for the confirming concrete example, and a median value of agree or better.

Qualitative Data. For the qualitative data, emerging codes were developed for Sect. 2 on a *per question* basis. This was so as to better interpret the quantitative results, and, in particular with the counter-examples, to see whether the reasons given by participants for their answers matched the theory behind the objections as given in the literature. Similar to the belief revision case, a common code was *new information should be believed*, which corresponds to the case of simply believing new information.

5 Results, Discussion and Conclusions

To answer our research question, we found several correspondences between the KLM [13] approach for defeasible reasoning, the AGM [1] approach for belief revision and the KM [11] approach for belief update, and how our participants reasoned. For defeasible reasoning, there was correspondence with one KLM [13] postulate (refer to Fig. 1): *Or*. For belief revision, there was correspondence with three AGM [1] postulates (refer to Fig. 2): *Success*, *Vacuity* and *Closure*. For belief update, there was correspondence with four KM [11] postulates (refer to

Fig. 3): $U1$, $U3$, $U4$ and $U6$. For each of the three surveys, we present additional results that are of importance. Our surveys were designed separately and contained slightly differing methodologies, so we have not attempted a holistic comparison of the results. Future work might do so. Discussion of less expected results from each survey can be found at either of the links in Appendix Sect. A.1, in the respective individual papers.

5.1 Additional Results for KLM Defeasible Reasoning

The KLM [13] postulate *Or* shows agreement with our participants, suggested by both the concrete and abstract hit rates being $\geq 50\%$ (concrete hit rate 75,86%, abstract hit rate 58,62%). In addition to *Or*, 2 out of 5 KLM [13] postulates show agreement in the concrete case only: *Left Logical Equivalence* (55,17%) and *Cautious Monotonicity* (72,41%). Across all KLM [13] postulates tested in this study, where both concrete and abstract hit rates were present, we observed the pattern that the concrete hit rate was always higher than abstract hit rate. The additional defeasible reasoning postulates we have investigated were *Rational Monotonicity*, *Transitivity* and *Contraposition*, included in Appendix A for reference. The defeasible reasoning postulate of *Transitivity* shows acceptance by our survey participants (concrete 72,41%, abstract 65,52%). In the case of *Contraposition*, a change in hit rate pattern was observed: it was the only postulate for which neither hit rate exceeded the baseline hit rate $\geq 50\%$, suggesting a negative relationship with our participants' reasoning, and it was the only postulate for which the abstract hit rate (41,38%) was higher than the concrete hit rate (17,24%). We observed that *Rational Monotonicity* had the largest difference between hit rates. The difference between the concrete and abstract hit rates for *Rational Monotonicity* was 65,51%, with a significant agreement in the concrete case (concrete 72,41%). In our investigation of participant agreement with prototypical reasoning and presumptive reasoning, we also observed strong agreement in the concrete case with both concrete hit rates $\geq 85\%$. Our additional investigation sought to identify whether participants reasoned normatively or descriptively. We found that across the majority of KLM [13] postulates and additional defeasible reasoning postulates, participants explained their acceptance or disagreement using a normative reasoning style. This can be explained by (1) participants relying mainly on the information given in the study and (2) participants accepting the information given in the study as plausible.

5.2 Additional Results for AGM Belief Revision

The hit rates were taken as indications of the type of relationship between human reasoning and the relevant AGM [1] postulates. For the concrete and abstract questions for postulates *Success* (concrete hit rate 90%, abstract hit rate 76.67%), *Closure* (concrete 100%, abstract 53.33%) and *Vacuity* (concrete 50%, abstract 56.67%), the hit rates obtained were $\geq 50\%$, suggesting a positive relationship between human reasoning and those postulates. Postulates *Extensionality* (concrete 26.67%, abstract 40%), *Super-expansion* (concrete 38.33%,

abstract 36.67%) and *Consistency* (concrete 50%, abstract 36.67%) received hit rates $\leq 50\%$, suggesting a negative relationship. Postulates *Sub-expansion* (concrete 76.67%, abstract 40%) and *Inclusion* (concrete 23.33%, abstract 60%) had discrepancies of $>30\%$ between the hit rates for their concrete and abstract questions, and their relationships to human reasoning thus found to be inconclusive. Through an additional investigation, we found that participants have a predominantly descriptive relationship with belief revision when postulates are presented both in concrete and abstract form. The balance of descriptive and normative reasoning styles of respondents in their responses became more even for the abstract questions, perhaps suggesting an increasing reliance on perceived rules in situations to which humans are less able to relate.

5.3 Additional Results for KM Belief Update

When reporting the abstract results in this section, the first number indicates the median of the participants' attitudes towards the postulate if 0 = strongly disagree and 5 = strongly agree. The subsequent percentage is the percentage of people who agreed or strongly agreed with the postulate. The concrete questions and counter-examples were sometimes tested using multiple questions. The hit rate for such postulates refers to the question with the lowest percentage of conformance with the postulate. *U1* (concrete 90%, abstract: 4; 76.7%), *U3* (concrete 76.7%, abstract: 4; 56.7%), *U4* (concrete 66.7%, abstract: 4; 76.7%) and *U6* (concrete 76.7%, abstract: 4; 66.7%) saw hit rates uniformly $>50\%$. *U2* (concrete 76.7%, abstract: 3; 50%) and *U5* (concrete 90%, abstract: 3; 50%) saw a neutral median abstract Likert score; with a correspondingly split abstract hit rate. *U7* (concrete 76.7%, abstract: 2.5; 36.7%) and *U8* (concrete 90%, abstract: 3; 43.3%) saw an abstract hit rate of $<50\%$. Qualitatively, all postulates excluding *U5* and *U7*, saw codes such that the majority reason for agreement with the concrete questions was theoretically in accordance with the postulate. All of the counter-examples examined saw hit rates $>50\%$. The first counter-example (hit rate 63.3%) followed from *U2*. The second counter-example (83.3%) follows from the set of *U1*, *U4*, and *U5*. The third counter-example (60%) was against *U8*. The final counter-example (70%) follows independently from *U4*, and the set of *U1* and *U6*. Although implicated theoretically, qualitative analysis suggested participants still reasoned in accordance with *U1* in the counter-examples.

6 Future Work

Our results suggest that the models of KLM defeasible reasoning [13], AGM belief revision [1] and KM belief update [11] are not yet a perfect fit with human reasoning because participants failed to reason in accordance with many of the postulates of these models. A larger participant pool is required to confirm our results. In future work, it may be interesting to add blocks to the study, in the form of different control groups e.g. paid reasoners as opposed to unpaid reasoners, to explore the effects of different circumstances on cognitive reasoning and which logic form is most closely resembled in each such block.

A Appendix: Supplementary Information

A.1 External Resources

We have created a GitHub repository which contains additional resources for this project. In this repository, we include our survey questions, our raw data and the codebooks used for our data analysis. As mentioned in the abstract, this work is the product of merging three independent papers: one each for KLM [13] defeasible reasoning, AGM belief revision [1] and KM belief update [11]. These independent papers are also included in the GitHub repository. The GitHub repository can be accessed [by clicking here](#). In addition, a summary of our project work is also showcased on our project website which can be viewed [by clicking here](#).

A.2 Defeasible Reasoning

KLM Postulates. Table 1 presents the KLM postulates. For ease of comparison, we present the postulates translated in a manner similar to [27]. We write $C_n(S)$ to represent the smallest set closed under classical consequence containing all sentences in S , and $D_C(S)$ to represent the resulting set if defeasible consequence is used instead. $D_C(S)$ is assumed defined only for finite S . $C_n(\alpha)$ is an abbreviation for $C_n(\{\alpha\})$, and $D_C(\alpha)$ is an abbreviation for $D_C(\{\alpha\})$.

Table 1. KLM postulates

1	Reflexivity	$\alpha \in D_C(\alpha)$
2	Left Logical Equivalence	If $\alpha \equiv \phi$ then $D_C(\alpha) = D_C(\phi)$
3	Right Weakening	If $\alpha \in D_C(\phi)$ and $\gamma \in C_n(\alpha)$ then $\gamma \in D_C(\phi)$
4	And	If $\alpha \in D_C(\phi)$ and $\gamma \in D_C(\phi)$ then $\alpha \wedge \gamma \in D_C(\phi)$
5	Or	If $\alpha \in D_C(\phi)$ and $\alpha \in D_C(\gamma)$ then $\alpha \in D_C(\phi \vee \gamma)$
6	Cautious Monotonicity	If $\alpha \in D_C(\phi)$ and $\gamma \in D_C(\phi)$ then $\gamma \in D_C(\phi \wedge \alpha)$

Reflexivity states that if a formula is satisfied, it follows that the formula can be a consequence of itself. *Left Logical Equivalence* states that logically equivalent formulas have the same consequences. *Right Weakening* expresses the fact that one should accept as plausible consequences all that is logically implied by what one thinks are plausible consequences. *And* expresses the fact that the conjunction of two plausible consequences is a plausible consequence. *Or* says that any formula that is, separately, a plausible consequence of two different formulas, should also be a plausible consequence of their disjunction. *Cautious Monotonicity* expresses the fact that learning a new fact, the truth of which could have been plausibly concluded, should not invalidate previous conclusions.

Additional Postulates. Table 2 presents additional defeasible reasoning postulates. *Cut* expresses the fact that one may, in his way towards a plausible

conclusion, first add an hypothesis to the facts he knows to be true and prove the plausibility of his conclusion from this enlarged set of facts and then deduce (plausibly) this added hypothesis from the facts. *Rational Monotonicity* expresses the fact that only additional information, the negation of which was expected, should force us to withdraw plausible conclusions previously drawn. *Transitivity* expresses that if the second fact is a plausible consequence of the first and the third fact is a plausible consequence of the second, then the third fact is also a plausible consequence of the first fact. *Contraposition* allows the converse of the original proposition to be inferred, by the negation of terms and changing their order.

Table 2. Additional postulates

1	Cut	If $\alpha \in D_C(\phi \wedge \gamma)$ and $\gamma \in D_C(\phi)$ then $\alpha \in D_C(\phi)$
2	Rational Monotonicity	If $\alpha \notin D_C(\phi \wedge \gamma)$ and $\neg\phi \notin D_C(\gamma)$ then $\alpha \notin D_C(\gamma)$
3	Transitivity	If $\alpha \in D_C(\phi)$ and $\gamma \in D_C(\alpha)$ then $\gamma \in D_C(\phi)$
4	Contraposition	If $\alpha \in D_C(\phi)$ then $\neg\phi \in D_C(\neg\alpha)$

A.3 Belief Revision

AGM Postulates. Table 3 presents the AGM postulates. $K * \alpha$ is the sentence representing the knowledge base after revising the knowledge base K with α . We assume that K is a set that is closed under classical deductive consequence.

Table 3. AGM postulates

1	Closure	$K * \alpha = C_n(K * \alpha)$
2	Success	$\alpha \in K * \alpha$
3	Inclusion	$K * \alpha \subseteq C_n(K \cup \{\alpha\})$
4	Vacuity	If $\neg\alpha \notin K$ then $C_n(K \cup \{\alpha\}) \subseteq K * \alpha$
5	Consistency	$K * \alpha = C_n(\alpha \wedge \neg\alpha)$ only if $\models \neg\alpha$
6	Extensionality	If $\alpha \equiv \phi$ then $K * \alpha = K * \phi$
7	Super-expansion	$K * (\alpha \wedge \phi) \subseteq C_n(K * \alpha \cup \{\phi\})$
8	Sub-expansion	If $\neg\phi \notin K$ then $C_n(K * \alpha \cup \{\phi\}) \subseteq K * (\alpha \wedge \phi)$

Closure implies logical omniscience on the part of the ideal agent or reasoner, including after revision of their belief set. *Success* expresses that the new information should always be part of the new belief set. *Inclusion* and *Vacuity* are motivated by the principle of minimum change. Together, they express that in the case of information α , consistent with belief set or knowledge base K , belief revision involves performing expansion on K by α i.e. none of the original beliefs need to be withdrawn. *Consistency* expresses that the agent should prioritise consistency, where the only acceptable case of not doing so is if the

new information, α , is inherently inconsistent - in which case, success overrules consistency. *Extensionality* effectively expresses that the content i.e. the belief represented, and not the syntax, affects the revision process, in that logically equivalent sentences or beliefs will cause logically equivalent changes to the belief set. *Super-expansion* and *sub-expansion* is motivated by the principle of minimal change. Together, they express that for two propositions α and ϕ , if in revising belief set K by α one obtains belief set K' consistent with ϕ , then to obtain the effect of revising K with $\alpha \wedge \phi$, simply perform expansion on K' with ϕ . In short, $K * (\alpha \wedge \phi) = (K * \alpha) + \phi$.

Table 4. KM postulates

1	(U1)	$\alpha \in K \diamond \alpha$
2	(U2)	If $\alpha \in K$ then $K \diamond \alpha = K$
3	(U3)	$K \diamond \alpha = C_n(\alpha \wedge \neg\alpha)$ only if $\models \neg\alpha$ or $K = C_n(\alpha \wedge \neg\alpha)$
4	(U4)	If $\alpha \equiv \phi$ then $K \diamond \alpha = K \diamond \phi$
5	(U5)	$K \diamond (\alpha \wedge \phi) \subseteq C_n(K \diamond \alpha \cup \{\phi\})$
6	(U6)	If $\phi \in K \diamond \alpha$ and $\alpha \in K \diamond \phi$ then $K \diamond \alpha = K \diamond \phi$
7	(U7)	If K is complete then $K \diamond (\phi \vee \alpha) \subseteq C_n(K \diamond \alpha \cup K \diamond \phi)$
8	(U8)	$K \diamond \alpha = \bigcap_{\phi \in K} C_n(\phi) \diamond \alpha$
9	(U*9)	$K \diamond \alpha = C_n(K \diamond \alpha)$

A.4 Belief Update

KM Postulates. Table 4 presents the KM postulates. For ease of comparison, the postulates have been rephrased as in the AGM paradigm [22]. We use \diamond to represent the update operator. *U1* states that updating with the new fact must ensure that the new fact is a consequence of the update. *U2* states that updating on a fact that could in principle be already known has no effect. *U3* states the reasonable requirement that we cannot lapse into impossibility unless we either start with it, or are directly confronted by it. *U4* requires that syntax is irrelevant to the results of an update. *U5* says that first updating on α then simply adding the new information γ is at least as strong (i.e. entails) as updating on the conjunction of α and γ . *U6* states that if updating on α_1 entails α_2 and if updating on α_2 entails α_1 , then the effect of updating on either is equivalent. *U7* applies only to complete knowledge bases, that is knowledge bases with a single model. If some situation arises from updating a complete K on α and it also results from updating that K from ϕ then it must also arise from updating that K on $\alpha \vee \phi$. *U8* is the disjunction rule. *U*9* is not necessary in the propositional formulation of the postulates and is listed for completeness. It was not tested in the survey.

A.5 Results

In Fig. 1, we show the Hit Rate (%) for each defeasible reasoning postulate. In Fig. 2, we show the Hit Rate (%) for each belief revision postulate. In Fig. 3, we show the Hit Rate (%) for each belief update postulate.

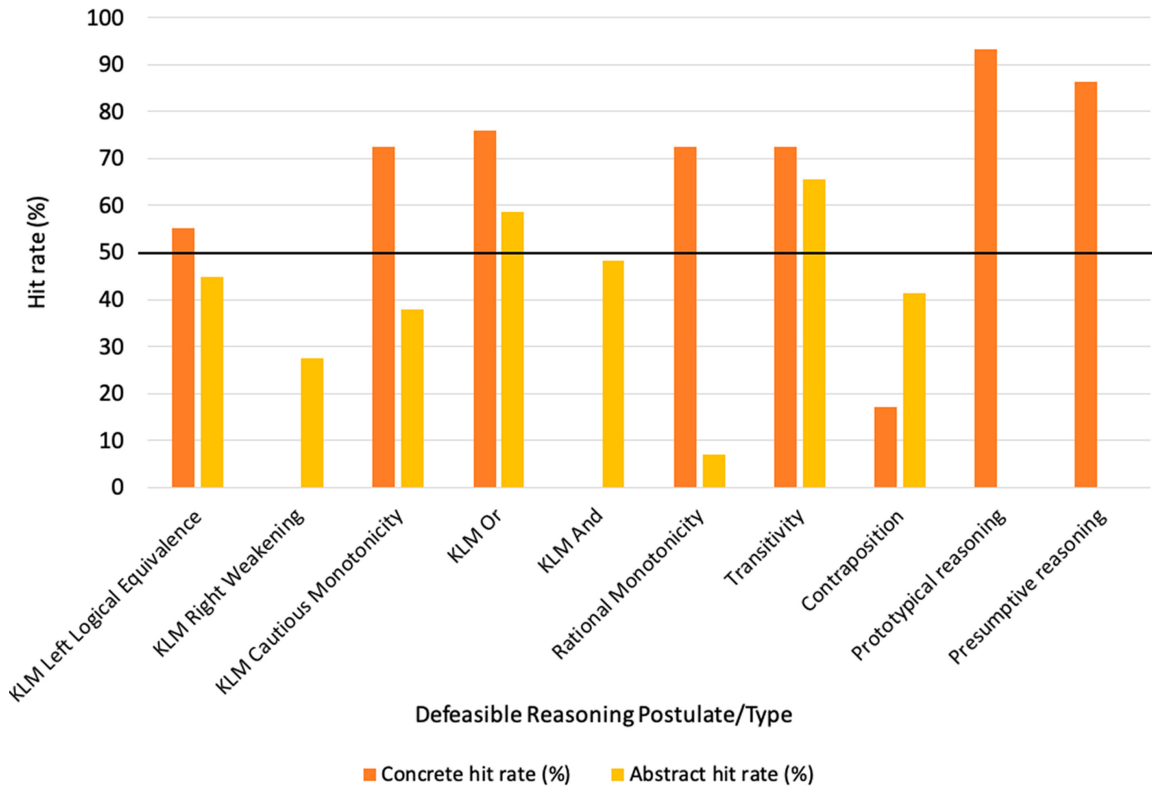


Fig. 1. Hit rate (%) for defeasible reasoning postulates

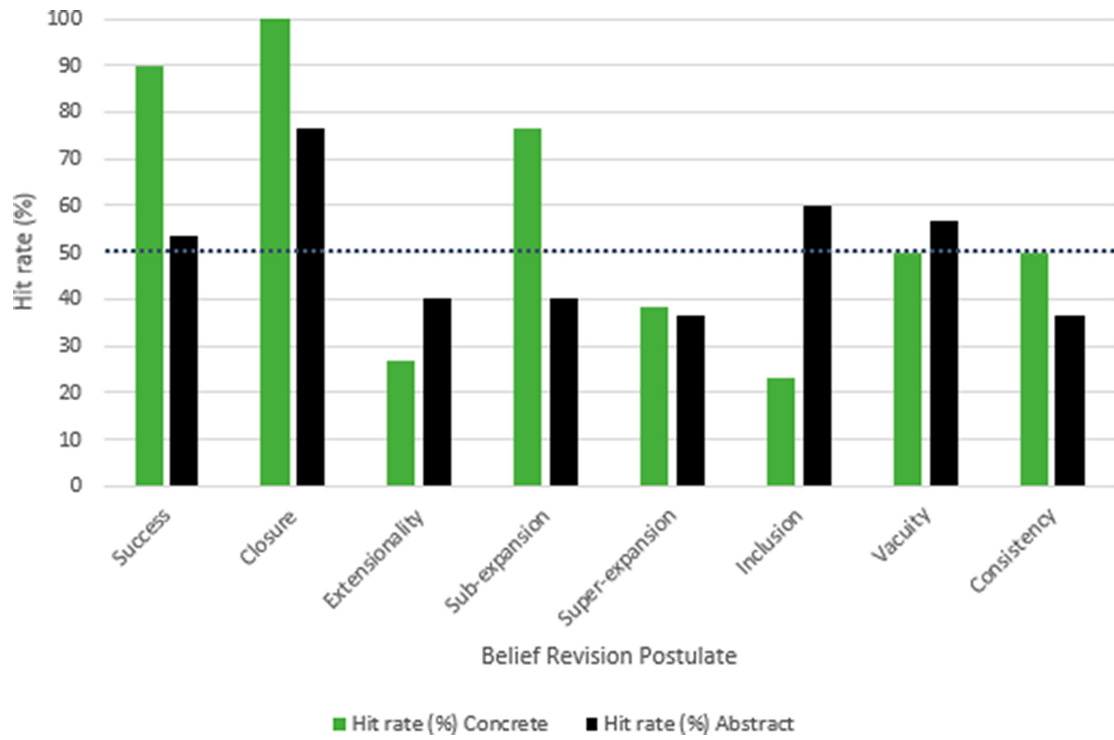


Fig. 2. Hit rate (%) for belief revision postulates

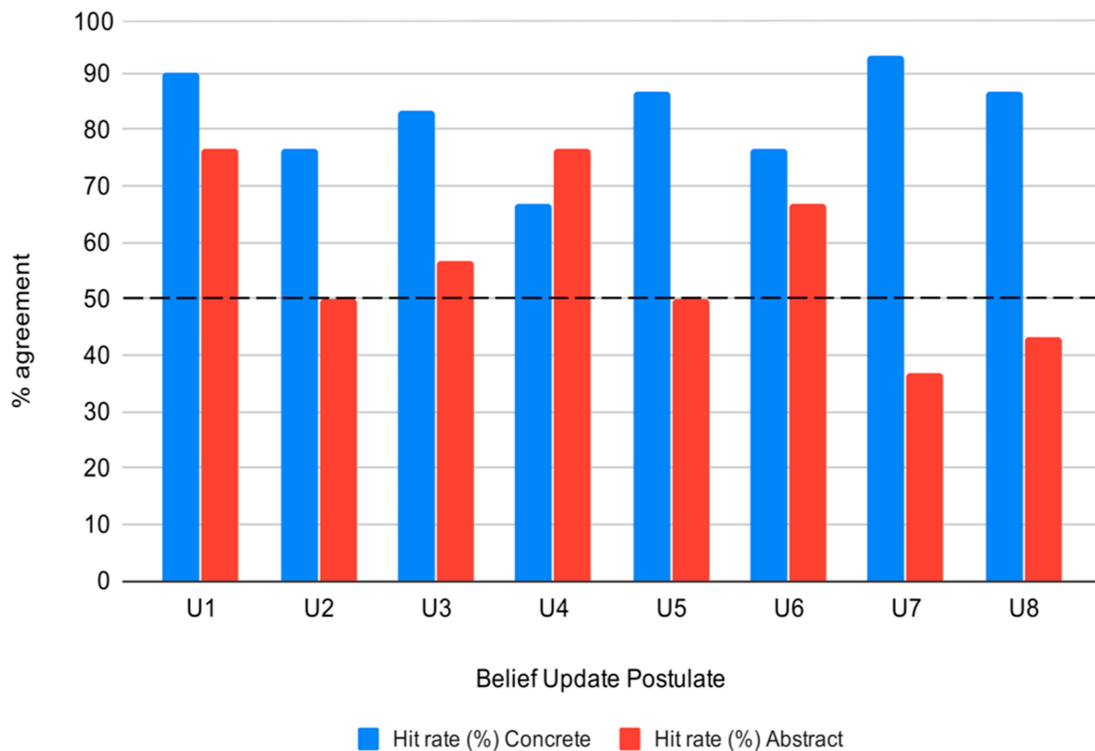


Fig. 3. Hit rate (%) for belief update postulates

References

1. Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: partial meet contraction and revision functions. *J. Symb. Logic* **50**, 510–530 (1985). <https://doi.org/10.2307/2274239>
2. Buhrmester, M.: M-turk guide (2018). <https://michaelbuhrmester.wordpress.com/mechanical-turk-guide/>
3. Creswell, J.W.: *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, vol. 4, pp. 245–253. SAGE Publications, Thousand Oaks (2014)
4. Darwiche, A., Pearl, J.: On the logic of iterated belief revision. *Artif. Intell.* **89**, 1–29 (1997). [https://doi.org/10.1016/S0004-3702\(96\)00038-0](https://doi.org/10.1016/S0004-3702(96)00038-0)
5. Gärdenfors, P., Makinson, D.: Nonmonotonic inference based on expectations. *Artif. Intell.* **65**(2), 197–245 (1994)
6. Gärdenfors, P.: *Belief Revision: An Introduction*, pp. 1–26. Cambridge University Press, Cambridge (1992). <https://doi.org/10.1017/CBO9780511526664.001>
7. Governatori, G., Terenziani, P.: Temporal extensions to defeasible logic. In: Orgun, M.A., Thornton, J. (eds.) *AI 2007. LNCS (LNAI)*, vol. 4830, pp. 476–485. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76928-6_49
8. Hansson, S.: *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Kluwer Academic Publishers, Berlin (1999)
9. Herzig, A., Rifi, O.: Update operations: a review. In: Prade, H. (ed.) *Proceedings of the 13th European Conference on Artificial Intelligence*, pp. 13–17. John Wiley & Sons, Ltd., New York (1998)
10. Inc., A.M.T.: *Faqs* (2018). <https://www.mturk.com/help>

11. Katsuno, H., Mendelzon, A.O.: On the difference between updating a knowledge base and revising it. In: Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning, KR 1991, pp. 387–394. Morgan Kaufmann Publishers Inc., San Francisco (1991). <http://dl.acm.org/citation.cfm?id=3087158.3087197>
12. Kennedy, R., Clifford, S., Burleigh, T., Jewell, R., Waggoner, P.: The shape of and solutions to the MTurk quality crisis, October 2018
13. Kraus, S., Lehmann, D., Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. *Artif. Intell.* **44**, 167–207 (1990)
14. Krosnich, J., Presser, S.: Question and questionnaire design. *Handbook of Survey Research*, March 2009
15. Lang, J.: Belief update revisited. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI 2007, pp. 1534–1540, 2517–2522. Morgan Kaufmann Publishers Inc., San Francisco (2007). <http://dl.acm.org/citation.cfm?id=1625275.1625681>
16. Lehmann, D.: Another perspective on default reasoning. *Ann. Math. Artif. Intell.* **15**(1), 61–82 (1995). <https://doi.org/10.1007/BF01535841>
17. Lieto, A., Minieri, A., Piana, A., Radicioni, D.: A knowledge-based system for prototypical reasoning. *Connect. Sci.* **27**(2), 137–152 (2015). <https://doi.org/10.1080/09540091.2014.956292>
18. Makinson, D.: Bridges between classical and nonmonotonic logic. *Logic J. IGPL* **11**(1), 69–96 (2003)
19. Martins, J., Shapiro, S.: A model for belief revision. *Artif. Intell.* **35**, 25–79 (1988). [https://doi.org/10.1016/0004-3702\(88\)90031-8](https://doi.org/10.1016/0004-3702(88)90031-8)
20. Over, D.: Rationality and the normative/descriptive distinction. In: Koehler, D.J., Harvey, N. (eds.) *Blackwell Handbook of Judgment and Decision Making*, pp. 3–18. Blackwell Publishing Ltd., United States (2004)
21. Pelletier, F., Elio, R.: The case for psychologism in default and inheritance reasoning. *Synthese* **146**, 7–35 (2005). <https://doi.org/10.1007/s11229-005-9063-z>
22. Peppas, P.: Belief revision. In: Harmelen, F., Lifschitz, V., Porter, B. (eds.) *Handbook of Knowledge Representation*. Elsevier Science, December 2008. [https://doi.org/10.1016/S1574-6526\(07\)03008-8](https://doi.org/10.1016/S1574-6526(07)03008-8)
23. Pollock, J.: A theory of defeasible reasoning. *Int. J. Intell. Syst.* **6**, 33–54 (1991)
24. Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., Tse, A.P.P.: Formal non-monotonic theories and properties of human defeasible reasoning. *Minds Mach.* **27**(1), 79–117 (2017). <https://doi.org/10.1007/s11023-016-9414-1>
25. Ragni, M., Eichhorn, C., Kern-Isberner, G.: Simulating human inferences in light of new information: a formal analysis. In: Kambhampati, S. (ed.) *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 16)*, pp. 2604–2610. IJCAI Press (2016)
26. Ross, J., Zaldivar, A., Irani, L., Tomlinson, B.: Who are the turkers? Worker demographics in Amazon mechanical turk, January 2009
27. Rott, H.: *Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning*. Oxford University Press (2001)
28. Sullivan, G., Artino, R., Artino, J.: Analyzing and interpreting data from likert-type scales. *J. Grad. Med. Educ.* **5**(4), 541–542 (2013)
29. Turk, A.M.: Qualifications and worker task quality best practices, April 2019. <https://blog.mturk.com/qualifications-and-worker-task-quality-best-practices-886f1f4e03fc>

30. TurkPrime: After the bot scare: Understanding what's been happening with data collection on mturk and how to stop it September 2018. <https://blog.turkprime.com/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it>
31. Verheij, B.: Correct grounded reasoning with presumptive arguments. In: Michael, L., Kakas, A. (eds.) JELIA 2016. LNCS (LNAI), vol. 10021, pp. 481–496. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48758-8_31
32. Witte, J.: Introduction to the special issue on web surveys. *Sociol. Methods Res.* **37**(3), 283–290 (2009)