Belief Change in Human Reasoning: An Empirical Investigation on MTurk*

Clayton Kevin Baker[1][0000−0002−3157−9989] and Thomas Meyer[1][0000−0003−2204−6969]

[1]University of Cape Town and Centre for Artificial Intelligence Research (CAIR), Cape Town, South Africa
clayton.baker@uct.ac.za
tmeyer@cair.org.za

Abstract. Belief revision and belief update are approaches to represent and reason with knowledge in artificial intelligence. Previous empirical studies have shown that human reasoning is consistent with non-monotonic logic and postulates of defeasible reasoning, belief revision and belief update. We extended previous work, which tested natural language translations of the postulates of defeasible reasoning, belief revision and belief update with human reasoners via surveys, in three respects. Firstly, we only tested postulates of belief revision and belief update, taking the position that belief change aligns more with human reasoning than non-monotonic defeasible reasoning. Secondly, we decomposed the postulates of revision and update into material implication statements of the form "If x is the case, then y is the case", each containing a premise and a conclusion, and then translated the premises and conclusions into natural language. Thirdly, we asked human participants to judge each component of the postulate for plausibility. In our analysis, we measured the strength of the association between the premises and the conclusion of each postulate. We used Possibility theory to determine whether the postulates hold with our participants in general. Our results showed that our participants' reasoning is consistent with postulates of belief revision and belief update when judging the premises and conclusion of the postulate separately.

Keywords: non-monotonic logic · AGM postulates · KM postulates · human reasoning · survey

1 Introduction

A non-monotonic inference is one that is well-drawn from given information, but may need to be withdrawn when we come into possession of new information, even when we do not discard the old premises [7]. Going further, Pollock [9]

* We wish to express our sincere gratitude and appreciation to the DSI – CSIR Interbursary Support (IBS) Programme and the Centre for Artificial Intelligence Research (CAIR) for financial support.

said that when one judges a situation based on how it looks to him, his reasoning is defeasible meaning that the premises alone may warrant accepting the conclusion, but when additional information is added, that conclusion may no longer be warranted. In the non-monotonic and defeasible case, it is clear that the conclusion of an argument based on some premises may change, should the premises change. In terms of reasoning with changing beliefs in artificial intelligence (AI), we apply non-monotonic reasoning to two unique cases: revising beliefs and updating beliefs. When we revise our beliefs, it means that we accept new information into our beliefs so long as the result of adding the information is consistent with our beliefs. Should a change in the world occur, due to some external phenomenon out of our control, revising our beliefs alone is insufficient. Instead, we can perform an update on our beliefs to indicate that we used to believe ψ, we learn that some event μ has occurred and thus we have to adapt our beliefs to include μ, whether or not μ contradicts ψ.

Empirical studies have shown that human reasoning and non-monotonic logic in AI share similarities. In the work of Da Silva Neves et al. [8], the translations of the KLM defeasible reasoning [6] postulates into English are tested in a survey and evaluated for plausibility, in the context of Possibility theory. Their approach involves translating the KLM postulates to rules of the form "if a then normally b" and using the Chi-square test to measure the significance of each component, the premises and the conclusion, of the postulates of preferential reasoning. They found that all except two postulates show conformance with their participants' reasoning. However, their analysis of how each component of the postulate relates to the evaluation of the overall postulate is not reproducible. Ragni et al. [10] have also motivated, with empirical evidence, that human reasoning exhibits conformance to the postulates of KLM defeasible reasoning. In previous work [2], we assessed the extent of correspondence between the postulates of KLM defeasible reasoning, AGM [1] belief revision and KM [5] belief update with human reasoning, via surveys conducted on Amazon's Mechanical Turk (MTurk, `https://www.mturk.com/`). Additionally, correspondence was evaluated through hit rates, where participants evaluated English translations of the postulates, on a linear Likert scale, ranging from strongly disagree to strongly agree. We found that this approach was too broad, however, and that correspondence between the postulates and human reasoning varied greatly across all three forms of reasoning. In this work, we refine our approach. We hypothesised that human belief change is consistent with the AGM postulates of belief revision and the KM postulates of belief update. We investigated this hypothesis in terms of the correspondence between human reasoning and the individual postulates (referred to as the postulate level) and in terms of the correspondence between human reasoning and the reasoning framework overall (referred to as the system level).

The outline of the rest of this paper is as follows. In Section 2, we describe our experimental setup and report on the experiment results. In Section 3, we discuss the results and how it answers our research hypothesis. We conclude our findings and present extensions to this work for future investigation in Sec-

tion 4. In Appendix A, we provide a URL to an external document, called "Paper_32_Supplementary_Material", which contains background material related to our investigation. In the external document, we also give a description of our methodology, ethical issues and additional results. Our project repository, containing experiment material, raw data, codebooks and scripts for data cleaning is accessible via the same URL. Finally, in Appendix B we provide a table with the URLs to all the surveys created for this project.

2 Experiments

Our investigation involved four experiments, with the task for each in the form of answering a survey. The first survey was designed in Microsoft Excel and responses were collected via email after holding a lottery on social media. For the remaining surveys, we used Google Forms (`https://www.google.com/forms`) for the design and used MTurk for data collection.

2.1 Experiment 1

In our first experiment, we prepared a survey of 30 general "if...then..." statements about the world for participants to evaluate for clarity and bias. An example of a general statement is: "If Jacob B is a truck driver then Jacob B does drive at night". The task was for participants to complete an evaluation table in which they identify statements whose language is unclear and statements which contain bias. Our sample consisted of 7 English-speaking participants who were recruited through social media to evaluate the general reasoning statements. Our participants were provided with instructions, an evaluation table and a statement table. To see the evaluation table, the complete instructions and survey material, visit our project repository via this URL, `https://tinyurl.com/y54epsmk`. Our results confirm all our statements were written in clear unambiguous English. In addition, as expected, our results also confirm the presence of two types of bias, global statement bias and local statement bias. Global statement bias (GBS) % refers to the percentage of participants who responded "Yes" to question (d) in the evaluation table. The GSB% is 85,71%. This means that 85,71% of participants expressed that at least one of statements 1 to 30 contained bias. Local statement bias (LSB)% for a statement S refers to the percentage of participants who both responded "Yes" to question (d) and specified a statement number S in the "Reason/Comment" column, where S corresponds to a particular statement in the survey material. The LSB% for 11 statements exceeded 50% and we concluded that those 11 statements contained bias. The LSB% for the remaining statements did not exceed the threshold and we concluded that those statements contained no bias. As a result, we have replaced the 11 biased statements with a mix of suggestions given by the survey participants and our knowledge.

2.2 Experiment 2

In our second experiment, we prepared a survey of 30 general statements about the world using the refined material obtained at the end of Experiment 1. The task was for participants to evaluate the degree to which they believe each of the statements in the survey, on a scale of 1 = strongly disagree to 5 = strongly agree, and provide an explanation for their answer. We suggested possible explanations that participants may have chosen to endorse. Participants were also given the option to provide their own explanation. Our sample consisted of 30 native English speakers recruited on MTurk with the following additional characteristics: the participant's *HIT Approval Rate (%) for all Requester's HITs* is greater than or equal to 98, the *Location* of all participants was restricted to the United States of America and the participant's *Number of HITs Approved* was greater than 50. Personal characteristics such as age, gender, race and education level were not pertinent to this experiment and were not collected. The survey itself was divided into 5 smaller surveys, each containing 6 statements to allow shorter survey response times and make the task less tedious for participants. The first survey we created on Google Forms, containing statements 1 to 6, can be accessed via this URL, `https://tinyurl.com/y7xh52us`. The second survey we created on Google Forms, containing statements 7 to 12, can be accessed via this URL, `https://tinyurl.com/2ptrw5ad`. The third survey we created on Google Forms, containing statements 13 to 18, can be accessed via this URL, `https://tinyurl.com/czfrc9yf`. The fourth survey we created on Google Forms, containing statements 19 to 24, can be accessed via this URL, `https://tinyurl.com/yt8pywy5`. The fifth survey we created on Google Forms, containing statements 25 to 30, can be accessed via this URL, `https://tinyurl.com/unvfyrx3`. We downloaded the responses from Google Forms and coded the data manually in Microsoft Excel. We considered the rank that a participant has given to a statement as qualitative data. We considered the explanation for the rank as qualitative data. We computed the frequency of each belief (number of 1s, 2s,...,5s) for each statement. In the next step of our analysis, we determined the average belief (AVB) for each statement using the formula: $\text{AVB} = \frac{\text{sum of individual beliefs}}{\text{total number of participants}}$. The AVB is a number representing the overall belief of a statement by our participants. We created a box plot of the average belief for each statement and show this in our supplementary material. A plausible statement is referred to as a statement from the general reasoning experiment which has an AVB greater than or equal to 4 and less than or equal to 5. According to this criteria, our participants found nine statements plausible and the remaining 21 statements implausible. In the survey, the provided explanations correspond to the interpretations of the statement in the form of a material implication rule. There are four interpretations for each statement, consisting of unique ways of assigning true and false to the premises ("if") and the conclusion ("then"). We computed the frequencies of the interpretations for each statement. In our supplementary material, we show the relative frequency (RF) % of explanation categories across all possible combinations of explanations. Overall, the 4 most frequently used explanations in order from first to fourth are A, B, D and

C. This shows that a majority of our participants (87,33%) preferred a single explanation over choosing multiple explanations and preferred a single explanation over providing their own explanations. After computing the frequencies, we determined the modal explanation category for each statement. From our data, we observed that there was a unique modal explanation category for each statement i.e. a unique explanation preferred by our participants as compared to all other explanations. Across the 30 statements, the modal explanation category was either A or B. This means that all our participants endorsed the premise of the 30 statements, but differed in the endorsement of the conclusion.

2.3 Experiment 3

In our third experiment, we prepared a survey of 18 rules corresponding to English translations of the 8 AGM postulates, obtained by first decomposing each postulate into its premises and conclusion, and then replacing each premise and conclusion with an even but random selection of plausible and implausible statements from Experiment 2. The task for participants was to rank the rules for plausibility on a scale of 1 = implausible to 10 = extremely plausible. We excluded tautological rules from our survey and assumed that our participants found tautological rules plausible. We recruited 50 participants on MTurk using the same criteria as Experiment 2. We divided our material over two shorter surveys to ease the mental load on participants. The survey testing the rules for the first four AGM postulates can be accessed via this URL, `https://tinyurl.com/wnw7aysy`. The survey testing the rules for the remaining four AGM postulates R5 to R8 can be accessed via this URL, `https://tinyurl.com/z523es9f`. Once collected, we coded our data in Microsoft Excel. We coded the plausibility ranking for each rule as either true or false, repeated for each participant. The code true was assigned to statements with a plausibility rating of greater than five and less than or equal to 10. The code false was assigned to statements with a plausibility rating of greater than or equal to 1 and less than 6. Once coded, rules were grouped according to the part of the postulate that it represents: the premises (LP) or the conclusion (RP). In the case of the premises of a postulate being a tautology, the overall logical valuation of the rules for the postulate translates to the overall valuation for the rules in RP. In our analysis, we wanted to determine whether the endorsement of LP (ELP) was preferentially associated with the endorsement of RP (ERP). This can be done in a contingency table with four cells, corresponding to combinations of endorsement of LP and RP. To identify the cell of the contingency table that a participant's response matched, we applied the logical evaluation of the implication LP $\longrightarrow$ RP. We computed the degree of association between ELP and ERP using the Phi-coefficient (ϕ), also known as the Matthews Correlation Coefficient (MCC), and tested the association statistically using Chi-square ($\mathcal{X}^2$). The value of ϕ for a 2 x 2 table is defined as:

$$\phi = \frac{|AD - BC|}{((A+B)(C+D)(A+C)(B+D))^{\frac{1}{2}}} \tag{1}$$

In the case of a zero denominator, where any of the four sums evaluate to zero, the denominator in the Phi-coefficient can be arbitrarily set to 1 [12,13], giving a Phi coefficient of 0. For each postulate, the statistical hypothesis H_0 (null hypothesis) was "there is no significant association between ELP and ERP". The alternate hypothesis, H_1, was "there is a significant positive association between ELP and ERP". H_0 was rejected if the probability of obtaining a value as large as the observed $\mathcal{X}^2$ was not greater than 0,05, as it is usual in experimental psychology [3, 4, 11]. Under our primary hypothesis that participants' inference tends to be consistent with postulates *R1*, *R2*, *R3*, *R4*, *R5*, *R6*, *R7* and *R8*, we predicted that H_0 would be rejected for each postulate. Under the hypothesis that participant judgements are consistent with the AGM belief revision framework, we predicted a high proportion of participants would not commit any postulate violations. Once the data was collected, a data cleaning process was performed. This step was programmed in R (`https://www.r-project.org/`) and the script is available in our project repository via this URL, `https://tinyurl.com/y54epsmk`. During the data collection process, 50 participants completed the first survey while only 35 participants completed both the first and second surveys. As a result, our adjusted sample size is 35 participants. MTurk uses the term *Workers* to refer to its participants who complete tasks. MTurk also uses the term *Requesters* to refer to users who create tasks for *Workers* to complete. An inspection of the data from MTurk revealed that the *Lifetime Approval Rate for Requesters Tasks (%) for all Workers* was 100%, on the last day of data collection, 29 April 2021. The number of tasks taken and approved ranged from 2 to 3. This indicates how familiar Workers were with our tasks on MTurk. 77,14% (27 Workers) have taken our tasks for the first time, the first and second parts of the survey for belief revision, with 22,86% having taken our tasks on a previous occasion, potentially having participated in our general reasoning survey, but potentially during research for a different project. An inspection of our data from Google Forms revealed that our 35 participants were divided by gender as 22 male (62,86%) and 13 female (37,14%), as shown in Figure 1. Participants' ages ranged from 22 to 74 years old. In Figure 2, we show the distribution of participant age by gender in a box plot. The box plot for male participants is comparatively short, which suggests that male participants were similar in age to the median of 40 years. The box plot for female participants is comparatively tall, which suggests that female participants were widely spread from the median age of 38 years. We show a box plot of the participant rank for our concrete rules in Figure 3. We observed that the box plot is comparatively short, spanning 2-3 consecutive ranks, for the majority of postulates. Additionally, all the rules with short box plots share the property that 75% of responses are greater than or equal to the rank of 6. Exceptions are the R5 premise rule, the R5 conclusion 2 rule, the R8 conclusion 1 rule and the R8 conclusion 2 rule, having comparatively tall box plots that span 4-5 consecutive ranks. Both exceptions come from postulates which involve multiple conclusions. This suggests that participants tend to differ on assigning a rank for rules which involve more than one conclusion. The box plots for the R1 conclusion 1 rule, the R1 conclusion 2 rule, the

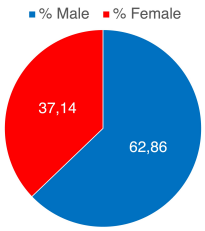

Fig. 1. Pie chart of participant gender distribution by percentage

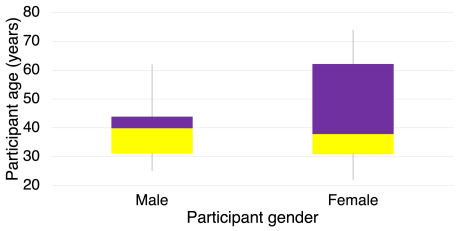

Fig. 2. Box plot of participant age distribution by gender

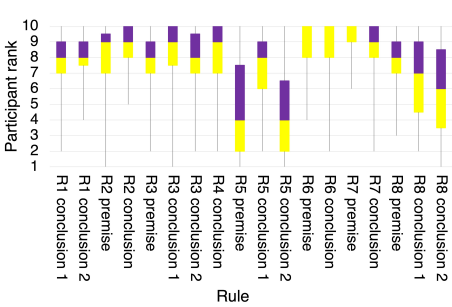

Fig. 3. Box plot of participant rule rank

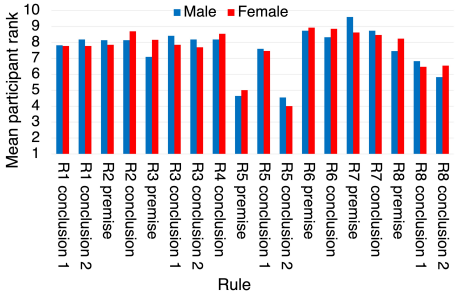

Fig. 4. Bar plot of average rule rank by gender

R3 premise rule, the R3 conclusion 2 rule, the R5 conclusion 1 rule and the R8 premise rule, share a median of 8, which suggests that there are few differences between the responses for those rules. Further, the R6 premise rule and the R6 conclusion rule share a median of 10, with 75% of responses greater than or equal to the rank of 8. Without further analysis, this suggests that the majority of our participants find AGM belief revision postulate R6 plausible. The box plot for the R5 premise rule and the R5 conclusion rule 2, share a median of 4 and a similar distribution in the third quartile, which is skewed to the right of the median. We show a bar plot of the average rank for each rule, coloured by gender in Figure 4. For each rule, the average rank was computed for male and female participants respectively. We note that according to this separation, our presentation of results in figure 4 is biased because male participants represent nearly two-thirds of our sample. However, an analysis of the average rank in this way is useful to identify whether our rule rank is influenced by gender. 16 rules have an average rank of 5 or above for both male and female participants, with two exceptions in the R5 premise rule and R5 conclusion 2 rule. Additionally, these exceptions show similar ranks between genders. The average rank between genders differs by at most 1-2 consecutive ranks for all rules. This suggests that overall our participants found our rules to be similarly plausible, regardless of gender. As part of our analysis, we determined the effect of the endorsement

of the premises and the endorsement of the conclusion of each postulate. We did this using contingency tables. We show the contingency tables, showing the spread of the endorsement of the premise and conclusion, for the AGM postulates in our supplementary material. Each contingency table consists of 2 rows and 2 columns and we refer to the top-left cell as A, the top-right cell as B, the bottom-left cell as C and the bottom-right cell as D. Each postulate has its contingency table except for postulates R1, which has been decomposed into two smaller postulates, each with its premise and conclusion rules. The values in cells C and D in the contingency tables for postulates with premises that are tautologies are 0 since these were not included in the survey. Across all postulates, strict violations - that is endorsements of LP but not of RP (B cells) - were only 8,89%.

$$\text{cell B endorsements} = \frac{\text{no. of B responses across all postulates}}{315} * 100 \qquad (2)$$

The denominator, 315, is equal to the product of the 35 responses obtained for each of the 9 postulates, counting R1 as two separate postulates. Endorsements of cells A, C and D were computed similarly. Endorsements of both LP and RP (A cells) were 78,41%. Non-endorsements of both RP and LP (D cells) were 6,98%. Non-endorsements of LP and endorsements of RP (C cells) were 5,71%. As a whole, the endorsement rates of LP were as expected. Focussing on the

Table 1. Values of ϕ and significance (p-value) of the association between LP and RP of each AGM belief revision postulate (N=35). "ns" means that ϕ is non-significant at the 0,05 level. "$*$" refers to the vanilla two-tailed Chi-square test."textdollar" refers to the two-tailed Chi-square test with Yates' correction. "†" refers to the two-tailed Fishers' Exact Test.

	ϕ	p-value $*$	p-value \$	p-value †
R1 (i)	0	-	-	ns
R1 (ii)	0	-	-	ns
R2	0,36	0,031	ns	ns
R3	0,12	ns	ns	ns
R4 (i)	0	-	-	ns
R5	0,09	ns	ns	ns
R6	0,8	$<$ 0,0001	0,0005	0,005
R7	0	-	-	ns
R8	0,4	0,0173	ns	ns

AGM belief revision postulates, the ϕ degrees of association reported in Table 1 show a highly significant association between ELP and ERP for postulate R6, using the two-tailed Chi-square test, Chi-square test with Yates' correction and Fisher's Exact Test. The ϕ degrees of association for postulates R2 and R8 are only significant using the two-tailed unmodified Chi-square test. Thus, there is evidence that the probability of mistakenly rejecting H_0 is lower than 0,005 for R2, R6 and R8, and is even lower than 0,0001 for R6. The ϕ degrees of association

reported in Table 1 also show a non-significant association between LP and RP for postulates R3 and R5. The significance between LP and RP for postulates R1 (i), R1 (ii), R4 (i) and R7 could not be determined because the conditions of the Chi-square test were not satisfied (zero-valued elements in cells C and D). As a consequence, we may reject the null hypothesis that there is no significant association between the premises and conclusion of R2, R6 and R8. Considering all postulates, we tallied how many violations (out of 9 possible) each participant made. Results appear in Table 2. 45,71% of participants made no violation at all, and 82,86% made one violation or none. These percentages are significantly high given the unavoidable imperfection of our apparatus, material, and participant sampling. These results suggest a pattern in which the majority of participants draw inferences from their beliefs in a manner which is consistent with the AGM belief revision postulates. In summary, our results have exhibited that postulates R2, R6 and R8 were significantly endorsed by our participants. The results for postulates R3 and R5 were non-significant, while the results for postulates R1 (i), R1(ii), R4(i) and R7 were inconclusive according to our criteria. At the system level, the majority of our participants committed one postulate violation or none.

Table 2. Percentages of participants as a function of the number of violations ($N = 35$)

Number of violations	0	1	2	3
Percentage of participants	45,71	37,14	8,57	8,57
Cumulative percentages of participants	45,71	82,86	91,43	100

Experiment 4 In our fourth experiment, we prepared a survey of 22 rules corresponding to English translations of the 9 KM postulates, obtained by decomposing each postulate into its premises and conclusion, and then replacing each component with an even but random selection of plausible and implausible statements from Experiment 2. The task for participants was the same as in Experiment 3. We excluded rules involving tautological premises in our survey material. We recruited 50 participants on MTurk using the same criteria as Experiments 2 and 3. The survey testing the rules for the first five KM postulates can be accessed via this URL, `https://tinyurl.com/xvufy32m`. The survey testing the rules for the remaining four KM postulates can be accessed via this URL, `https://tinyurl.com/3sxs9rjz`. During the data collection process, 50 participants completed the first survey while only 37 completed both the first and second surveys. As a result, our adjusted sample size is 37 participants. Once collected, a data cleaning step was performed. This step was programmed in R and the script is available in our project repository via this URL, `https://tinyurl.com/y54epsmk`. We coded our data in Microsoft Excel following the same procedure as Experiment 3. An inspection of the data from

MTurk revealed that the *Lifetime Approval Rate for Requesters Tasks (%) for all Workers* was 100%, on the last day of data collection, 5 July 2021. The number of tasks taken and approved ranged from 2 to 5. This indicates how familiar Workers are with our tasks on MTurk. 81,08% (30 Workers) have taken our tasks for the first time, the first and second parts of the survey for belief update, with 8,11% having taken 1 of our tasks previously, an additional 8,11% having taken 2 of our tasks previously and a further 2,7% having taken 3 of our tasks previously. An inspection of our data from Google Forms revealed that our 37 participants were divided by gender as 19 male (51,35%) and 18 female (48,65%), as shown in Figure 5. Participants' ages ranged from 24 to 61 years old. In Figure 6, we show the distribution of participant age by gender in a box plot. The box plot for male participants has a similar height to the box plot for female participants. This suggests that the age range for male and female participants was similar. However, the distribution for male and female participants is different in quartiles 2 and 3: males participants are skewed to the left of their median of 37, while female participants are distributed evenly around their median of 43. This suggests that the female participants were more similar in age to the median than their male counterparts. The box plot for female participants is higher than the box plot for male participants. This suggests that female participants tended to be older than male participants in our study. We show a box plot of

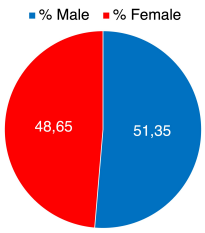

Fig. 5. Pie chart of participant gender distribution by percentage

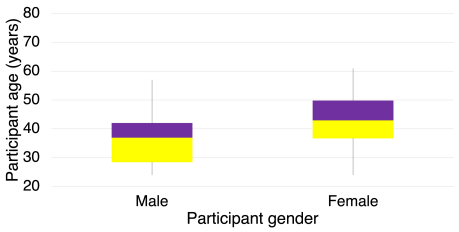

Fig. 6. Box plot of participant age distribution by gender

the participant rank for our 22 concrete rules in Figure 7. We observed that the box plot is comparatively short, spanning 2-3 consecutive ranks, for the majority of postulates. Additionally, all the rules with short box plots share the property that 75% of responses are greater than or equal to the rank of 6. Exceptions are the U2 premise rule, the U4 premise 1 rule, U4 premise 2 rule, U4 premise 3 rule, U4 premise 4 rule, U4 conclusion 2 rule, U7 premise rule and U8 conclusion 2 rule, having comparatively tall box plots that span 4-5 consecutive ranks. The exceptions occur in all of the premise rules of postulates U2, U4 and U7, and in one of the conclusion rules of both postulate U4 and postulate U8. Postulate U4 is the only postulate where the variance in the rank spans 4-5 consecutive ranks in both the premise and conclusion rules. This suggests that U4 is a contentious postulate with ranks ranging from moderate to high. In turn, the variance in the

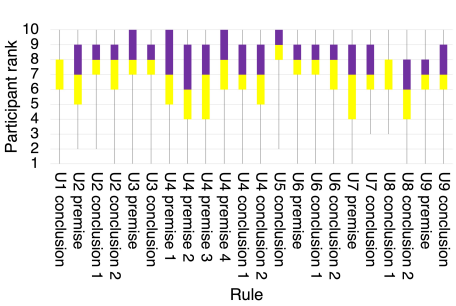

Fig. 7. Box plot of participant rule rank

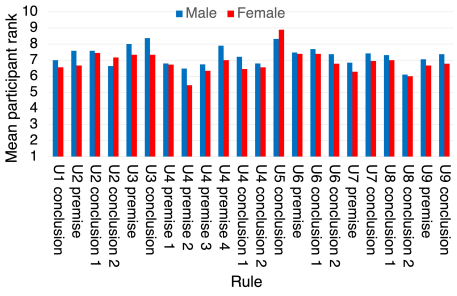

Fig. 8. Bar plot of average rule rank by gender

rank of the premises of postulates U2, U4 and U7, suggest that those postulates have contentious premises with ranks also ranging from low-moderate to high. The box plots for the U1 conclusion rule, U2 conclusion 1 rule, U2 conclusion 2 rule, U3 premise rule, U3 conclusion rule, U4 premise 4 rule, U6 premise rule, U6 conclusion rule, U6 conclusion 2 rule, U8 conclusion 1 rule and U9 premise rule, share a median rank of 8. This suggests that there are few differences between the responses for those rules. Additionally, the U5 conclusion rule has a median rank of 9 with 75% of responses greater than or equal to the rank of 8. Without further analysis, our results suggest that the majority of our participants find belief update postulate U4 contentious while our participants seem to find postulates U6 and U9 plausible. We show a bar plot of the average rank for each rule, coloured by gender in Figure 8. For each rule, the average rank was computed for male and female participants respectively. We note that according to this separation, our presentation of results in figure 8 should not be biased because the number of male and female participants only differ by 1 member. In our analysis, we determined whether our rule rank was influenced by gender. All rules have an average rank of 6 or above for both male and female participants, with one exception in the U4 premise 2 rule. Additionally, this exception differed for male and female participants by only one rank. The average rank between genders differed by at most 1-2 consecutive ranks for all rules. This suggests that overall our participants found our rules to be similarly plausible, regardless of gender. Our data analysis for this experiment follows the same steps as for the belief revision experiment. We show the contingency tables for the KM belief update postulates in our supplementary material. Each postulate has its own contingency table. The values in cells C and D in the contingency tables for postulates with premises that are tautologies are 0 since these were not included in the survey. Considering all postulates, strict violations - defined as the endorsements of LP but not of RP (B cells) - were only 12,61%.

$$\text{cell B endorsements} = \frac{\text{no. of B responses across all postulates}}{333} * 100 \qquad (3)$$

The denominator, 333, is equal to the product of the 37 responses obtained for each of the 9 postulates, U1 to U9. Endorsements of cells A, C and D were computed similarly. Endorsements of both LP and RP (A cells) were 66,67%. Non-endorsements of both RP and LP (D cells) were 12,91%. Non-endorsements of LP and endorsements of RP (C cells) were 7,81%. As a whole, the endorsement rates of LP were as expected. Focussing on the KM belief update postulates, the

Table 3. Values of ϕ and significance (p-value) of the association between LP and RP of each KM belief update postulate (N=37). "ns" means that ϕ is non-significant at the 0,05 level. "$*$" refers to the vanilla two-tailed Chi-square test."$" refers to the two-tailed Chi-square test with Yates' correction. "†" refers to the two-tailed Fisher's Exact Test.

	ϕ	p-value $*$	p-value $	p-value †
U1	0	-	-	ns
U2	0,66	$<$ 0,0001	0,0003	0,0002
U3	0,54	0,0009	0,0204	0,0225
U4	0,68	$<$ 0,0001	0,0002	$<$ 0,0001
U5	0	-	-	ns
U6	0,71	$<$ 0,0001	$<$ 0,0001	$<$ 0,0001
U7	0,28	0,0885	ns	ns
U8	0	-	-	ns
U9	0,31	0,0559	ns	0,0784

ϕ degrees of association reported in Table 3 show a highly significant association between LP and RP for postulates U2, U4 and U6, and a weakly significant association between LP and RP for postulate U3, using the two-tailed Chi-square test, Chi-square test with Yates' correction and Fisher's Exact Test. The ϕ degrees of association for postulates U7 and U9 are only significant using the two-tailed unmodified Chi-square test. Thus, there is evidence that the probability of mistakenly rejecting H_0 is lower than 0,005 for U3, U7 and U9 and is even lower than 0,0001 for U2, U4 and U6. The ϕ degrees of association reported in Table 3 also show a non-significant association between LP and RP for postulates U1, U5 and U8, using Fisher's Exact Test. Postulate U7 has a non-significant association between LP and RP in the cases of the Chi-square test with Yates' correction and Fisher's Exact Test, but not the vanilla Chi-square test. U9 has a non-significant association between LP and RP only in the case of the Chi-square test with Yates' correction. The significance between LP and RP for postulates U1, U5 and U8, using the vanilla Chi-square test, could not be determined because the conditions of the Chi-square test were not satisfied (zero-valued elements in cells C and D). As a consequence, we may reject the null hypothesis that there is no significant association between the premises and conclusion of U1, U5 and U8. Considering all postulates, we tallied how many violations (out of 9 possible) each participant made. Results appear in Table 4. 18,92% of participants made no violation at all, and 70,27% made one violation or none. These percentages are significantly high given the unavoidable imperfection of our experimental

apparatus, material, and participant sampling. These results suggest a pattern in which the majority of participants draw inferences from their beliefs in a manner which is consistent with the KM belief update postulates. In summary, our results have exhibited that postulates U2, U3, R4, U6, U7 and U9 were significantly endorsed by our participants. The results for postulates U1, U5 and U8 were inconclusive, according to our criteria. At the system level, the majority of participants committed one violation or none.

Table 4. Percentages of participants as a function of the number of violations ($N = 35$)

Number of violations	0	1	2	3
Percentage of participants	18,92	51,35	27,03	2,7
Cumulative percentages of participants	18,92	70,27	97,3	100

3 Results and discussion

In this section, we discuss the results from the belief revision experiment and the belief update experiment. We discuss how these results relate to and answer our research hypotheses, which states human belief change is consistent with AGM postulates of belief revision, and which states human belief change is consistent with the KM postulates of belief update.

3.1 Human reasoning and the AGM postulates

We presented evidence from our experiments that our participants found our concrete instantiations of the AGM belief revision postulates plausible and then we determined whether the postulates hold in general. Our results show that our participants' reasoning tends to be consistent with the 9 AGM belief revision postulates, with the significance of the association between the endorsement of the premises and the endorsement of the conclusion for each postulate ranging from non-significant to highly significant. Of the 9 postulates, the majority show a low number of violations (less than 50% of responses located in cell B for each postulate). Additionally, our hypothesis that our participants' reasoning is consistent with the AGM postulates is proven true for 3 postulates: R2, R6 and R8. Our hypothesis that our participants' reasoning is consistent with the AGM postulates is determined to be inconclusive for 6 postulates: R1 (i), R1 (ii), R3, R4, R5 and R7. In the case of postulates R1 (i), R1 (ii) and R4, the general form of the postulate resembled the following rule: $\top \rightarrow q$. For these postulates, our specific concrete rules used to instantiate the postulates were found plausible by our participants. However, due to the statistically non-significant association between the endorsement of the premises and the endorsement of the conclusion

in all 6 cases, we could not conclude that our participants' reasoning was consistent with all rules of the form $\top \rightarrow q$. We propose an additional investigation is needed to assess whether all rules of the form $\top \rightarrow q$ are found plausible by our participants and suggest the inclusion of the tautological premises as a rule in the survey for participants to rank as a starting point for future analyses. Additionally, postulates R2 and R6 share the same rule form: $p \rightarrow q$. In both cases, the association between the endorsement of the premises and the endorsement of the conclusion is statistically significant to highly statistically significant. This suggests a strong assertion by our participants that rules of the form $p \rightarrow q$ are generally plausible. In the case of postulate R8, the general rule is described by $p \rightarrow (q_1 \wedge q_2)$, where p represents a single premises and $q_1 \wedge q_2$ represents a conjunction of two conclusions. The association between the endorsement of the premises and the endorsement of the conclusion of postulate R8 is statistically significant. This suggests a strong assertion by our participants that rules of the form $p \rightarrow (q_1 \wedge q_2)$ are generally plausible.

3.2 Human reasoning and the KM postulates

We presented evidence from our experiments that our participants found our concrete instantiations of the KM belief update postulates plausible and then we determined whether the postulates hold in general. Our results show that our participants' reasoning tends to be consistent with the 9 postulates of KM belief update, with the significance of the association between the endorsement of the premises and the endorsement of the conclusion of each postulate ranging from non-significant to highly significant. Of the 9 postulates, the majority committed a low number of violations (less than 50% of responses located in cell B for each postulate) with no participants committing a violation in the case of postulate U4. Our hypothesis that our participants' reasoning is consistent with the KM postulates of belief update is proven true for 4 postulates: U2, U3, U4 and U6. For these postulates both the specific concrete rules used to instantiate the postulates and the general form of the rule was found plausible by our participants. The general rule for postulate U2 is given by: $p \rightarrow (q_1 \wedge q_2)$. This form is shared with postulate U6. The general rule for postulate U3 is given by: $p \rightarrow q$. The general rule form for postulate U4 is given by: $(p_1 \wedge p_2 \wedge p_3 \wedge p4) \rightarrow (q_1 \wedge q_2)$. Our hypothesis that our participants' reasoning is consistent with the KM postulates of belief update was inconclusive for 6 postulates: U1, U5, U7, U8 and U9. For these postulates, the specific concrete rules used to instantiate the postulates were found plausible by our participants. However, due to the statistically non-significant association between the endorsement of the premises and the endorsement of the conclusion in all 5 cases, we could not conclude that our participants' reasoning was consistent with all rules of the form $\top \rightarrow q$, in the case of postulates U1 and U5, a rule of the form $p \rightarrow q$ in the case of postulates U7 and U9, and a rule of the form $p \rightarrow q_1 \wedge q_2$ in the case of postulate U8. Some of our results for the KM belief update postulates are contradictory. An example of this is postulate U3, U7 and U9 that all share the general form of $p \rightarrow q$. From the experiments, we found evidence for our participants' reasoning being

consistent with postulate U3, but neither postulate U7 nor postulate U9. In the survey testing postulate U3, we used and defined the term "satisfiable". In the survey testing postulates U7 and U9, we used and defined the term "complete". We suggest that our participants wrestled with the meaning of complete beliefs and complete belief bases. This is not unexpected as our participants are not experts in formal logic. Further, the concrete rules used in both U7 and U9 involve complex conjunction statements which may further have confused our participants. We propose an additional investigation is needed to assess whether all rules of the form $\top \rightarrow q$ and $p \rightarrow q_1 \wedge q_2$ are found plausible.

4 Conclusions and future work

Based on previous empirical studies with human reasoners and the link between human reasoning and belief change in AI, we hypothesised that human belief change is consistent with the AGM postulates of belief revision and the KM postulates of belief update. We investigated this hypothesis through four experiments, with the task for each in the form of answering a survey. For the first experiment, we prepared a survey of 30 general statements about the world, for example, "If Jacob B is a truck driver then Jacob B does drive at night". 7 English-speaking participants were recruited via a lottery held on social media and the task was to evaluate the statements for clarity and bias. Overall, 11 statements contained bias and we replaced these statements with a mix of suggestions given by the survey participants and our knowledge, for use in the remaining experiments. For the second experiment, we prepared a survey of 30 general statements about the world using the refined material from Experiment 1. 30 English-speaking participants were recruited from MTurk for this experiment. The task was for participants to evaluate the degree to which they believe each of the statements in the survey, on a scale of 1 = strongly disagree to 5 = strongly agree, and provide an explanation for their answer. We suggested explanations for participants to endorse and allowed them to provide their own. Our results show that participants found 9 statements plausible while the remaining statements were found implausible. Additionally, participant explanations were analysed and we found that participants preferred endorsing a single explanation over endorsing multiple explanations or providing their explanations. Our analysis shows that a modal explanation category exists for each statement and that participants were in agreement on the endorsement of the premises of the statements in the survey, but differed on the endorsement of the conclusions of the statements. For the third and fourth experiments, we recruited 50 English-speaking participants on MTurk to rank, on a scale from 1 = implausible to 10 = extremely plausible, the plausibility of the English translations of the AGM and KM postulates when formulated as material implication statements. We measured the association between the premises and conclusion of each postulate and tested the association statistically using Chi-square, Chi-square with Yates' correction and Fisher's Exact Test. We used Possibility theory to determine whether our participants' reasoning is consistent with the AGM and KM postulates in

general. Of the 9 AGM postulates, the majority committed a low number of violations by endorsing the premises of a postulate but not its conclusion. We presented evidence for AGM postulates R2, R6 and R8 being consistent with our participants' reasoning, while the findings for AGM postulates R1(i), R1(ii), R3, R4, R5 and R7 were inconclusive. Similarly, the results from the KM experiment a low number of postulate violations. We presented evidence for KM postulates U2, U3, U4 and U6 being consistent with our participants' reasoning, while the findings for KM postulates U1, U5, U7, U8 and U9 were inconclusive. Some of our results from the KM experiment are contradictory, for example, in the case of postulates U3, U7 and U9 which share the same rule form, $p \rightarrow q$, with postulate U3 found plausible by our participants, but not postulates U7 and U9. We propose that an additional investigation is needed to evaluate (i) whether all material implication rules with tautological premises and (ii) whether all rules of the form $\top \rightarrow q$ and $p \rightarrow q_1 \wedge q_2$ are plausible in human belief change. We submit that the translation into English and the decomposition of the AGM and KM postulates, into its premises and conclusion, is not straightforward. As a result, we have demonstrated that the representation of the postulates does have a statistically significant influence in determining whether the postulates are consistent with human reasoning.

References

1. Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: Partial meet contraction and revision functions. Journal of Symbolic Logic **50**, 510–530 (1985). https://doi.org/10.2307/2274239 2
2. Baker, C., Denny, C., Freund, P., Meyer, T.: Cognitive defeasible reasoning: the extent to which forms of defeasible reasoning correspond with human reasoning. In: Proceedings of the First Southern African Conference for Artificial Intelligence Research (SACAIR 2020). CCIS, Springer (2020) 2
3. Hentschke, H., Stüttgen, M.: Computation of measures of effect size for neuroscience data sets. European Journal of Neuroscience **34**(12), 1887–1894 (2011) 6
4. Johnson, V., Payne, R., Wang, T., Asher, A., Mandal, S.: On the reproducibility of psychological science. Journal of the American Statistical Association **112**(517), 1–10 (2017) 6
5. Katsuno, H., Mendelzon, A.O.: On the difference between updating a knowledge base and revising it. Belief revision p. 183 (1991) 2
6. Kraus, S., Lehmann, D., Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence **44**, 167–207 (1990) 2
7. Makinson, D.: Bridges from classical to nonmonotonic logic. Texts in computing ; v. 5, King's College, London (2005) 1
8. Neves, R., Bonnefon, J., Raufaste, E.: An empirical test of patterns for nonmonotonic inference. Annals of Mathematics and Artificial Intelligence **34**(1-3), 107–130 (2002) 2
9. Pollock, J.: Defeasible reasoning. Cognitive Science **11**(4), 481–518 (1987). https://doi.org/https://doi.org/10.1016/S0364-0213(87)80017-4, http://www.sciencedirect.com/science/article/pii/S0364021387800174 1

10. Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., Tse, A.: Formal nonmonotonic theories and properties of human defeasible reasoning. Minds and Machines **27**, 79–117 (2017). https://doi.org/10.1007/s11023-016-9414-1 2
11. Rana, R., Singhal, R.: Chi-square test and its application in hypothesis testing. Journal of the Practice of Cardiovascular Sciences **1**(1), 69 (2015) 6
12. Souza, E., Negri, T.: First prospects in a new approach for structure monitoring from gps multipath effect and wavelet spectrum. Advances in Space Research **59**(10), 2536–2547 (2017) 6
13. Zhao, F., Xu, J., Lin, Y.: Similarity measure for patients via a siamese cnn network. In: Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing. pp. 319–328. Springer (November 2018) 6

A Supplementary Material

We have created an external document, called "Paper_32_Supplementary_Material", with background material for our investigation. In the document, we also give a description of our methodology, ethical issues and additional results. This document can be accessed in our project repository on Github via this URL, `https://tinyurl.com/y54epsmk`. Additionally, our survey material, keyword dictionary, code scripts and related artefacts can be accessed via the same URL.

B Surveys

For convenience, the URLs to our surveys are included in Table 5.

Table 5. Survey URLS

Experiment no.	Survey description	URL
1.	Evaluation form	`https://tinyurl.com/2dhntz2w`
2.	Statements 1 to 6 of 30	`https://tinyurl.com/y7xh52us`
	Statements 7 to 12 of 30	`https://tinyurl.com/2ptrw5ad`
	Statements 13 to 18 of 30	`https://tinyurl.com/czfrc9yf`
	Statements 19 to 24 of 30	`https://tinyurl.com/yt8pywy5`
	Statements 25 to 30 of 30	`https://tinyurl.com/unvfyrx3`
3.	AGM Postulates R1 to R4	`https://tinyurl.com/wnw7aysy`
	AGM Postulates R5 to R8	`https://tinyurl.com/z523es9f`
4.	KM Postulates U1 to U5	`https://tinyurl.com/xvufy32m`
	KM Postulates U6 to U9	`https://tinyurl.com/3sxs9rjz`