An Empirical Investigation for Patterns of Belief Change in Human Reasoning

by

Clayton Kevin Baker

A dissertation presented for the degree of
Master of Science
in the
Department of Computer Science, University of Cape Town
February 2022
Supervisor: Prof. Thomas A. Meyer

Abstract

Empirical methods have been used to test whether human reasoning conforms to models of reasoning in logic-based artificial intelligence. Particularly, studies have shown that human reasoning is consistent with non-monotonic logic and belief change. The former refers to models of reasoning where previously drawn conclusions can be retracted when adding new information. The latter refers to the operations of revising and updating a set of beliefs, respectively, when presented with new information. The operation of revision assumes the world has remained the same and that only our knowledge about the world has changed. On the other side, update assumes that the world may have changed by the time the new information is added. Both operations require the resulting belief set to be consistent. Revision allows inconsistent beliefs to be deleted, while update acknowledges that the belief set may have been wrong if it contradicts the new information. Our work surveyed postulates of belief change with human reasoners. First, we studied the role of the postulates of belief revision and belief update in the literature. Next, we decomposed the postulates of revision and update into material implication statements, each containing a premise and a conclusion. We translated the premises and conclusions into English and surveyed the postulate translations with human reasoners. The main task of the surveys was for participants to judge the translated postulate components for plausibility. For our data analysis, we used statistical methods to test the relationship between the endorsement of the premises and the endorsement of the conclusion. For our data validation, we applied possibility theory to examine whether the relationship was significant for the broader population of English-speaking human reasoners. The results show that our participants' reasoning tends to be consistent with the postulates of belief revision and belief update when judging the premises and conclusion of the postulate separately.

Dedication

To my grandmother, Ma Julia Christina Robertson (née Parrott), who inspired me to pursue higher education and who took great joy in hearing about my success.
To mum, Priscilla Gloria Lorraine Baker (née Robertson), for her unwavering support and love.
To my younger brother, Gareth Ryan Baker, for his friendship and humour.
Lastly, but by no means least, to my extended family, relatives, friends and colleagues who have experienced hardship and loss during the Covid-19 pandemic.

Declaration

I know the meaning of plagiarism and declare that all work in the dissertation, "An Empirical Investigation for Patterns of Belief Change in Human Reasoning", save for that which is properly acknowledged, is my own.

Signed: Clayton K. Baker

Date: 09 February 2022

Acknowledgements

I wish to thank my Master's supervisor, Prof. Tommie Meyer, for his confidence, guidance and wisdom throughout this research project.
I wish to express my sincere gratitude and appreciation to the DSI — CSIR Interbursary Support (IBS) Programme and the Centre for Artificial Intelligence Research (CAIR) for financial support.

Contents

Chapter 1

Introduction

Artificial Intelligence (AI) is a field of study in which intelligent behaviour is realised through computation. From Russell and Norvig's [59] exposition, it is apparent that AI has divergent goals. One is to realise natural intelligence as observed in humans. The other is to realise perfect intelligence according to a standard of rationality. These goals have a common thread, however, in that both require a form of capturing knowledge and a capacity to make inferences from it. A unique field exists in AI to deal with this problem. Brachman et al. [17] term this field as knowledge representation and reasoning. Through the use of logic, the study of entailment relations, we are able to represent what we know and make inferences from it. Logic is important for argument formation, modelling complex syntax in natural language and judging the validity of an argument. For example, it is essential to distinguish between "Andy will graduate from university if he passes his final exams" and "Andy will graduate university if and only if he passes his final exams" if we want to reason about Andy's chances for graduating from university. Many forms of logic exist. Some examples are deductive logic, inductive logic, modal logic, propositional logic and first-order logic. While Van Harmelen et al. [63] express that first-order logic is extensively used in the theory of knowledge representation, there is no one-size-fits-all approach. The choice of logic determines the level of expressivity and complexity of the language. For example, classical logic deals with well-defined formulas concretely by assigning fixed truth values to each instance. It has the property of monotonicity, that once a conclusion is reached then the conclusion will remain the same even if we add some information to our knowledge or refine our knowledge. In practice, human reasoning does not follow the monotonic-

ity property. This led to the study of non-monotonic logics, which allows for conclusions to change under certain conditions. The hallmark non-monotonic reasoning problem in AI is that of "Tweety the bird" which presents the evidence that penguins are non-flying birds and that this evidence defeats the accepted premise that birds can fly [53]. In the cognitive science and AI communities, an ongoing goal is to model the way people draw conclusions in everyday situations [60]. Ragni et al. [54] argue that future technologies will increasingly interact with humans and demonstrate cognitive features such as tolerance to exceptions, robustness, and the flexibility to accommodate new information.

1.1 Problem Statement

Non-monotonic logic is an approach to reasoning in which the same set of premises does not always yield the same conclusion. This has been shown to hold in human reasoning to some extent, but the methodologies that test this relationship differ within the AI community. An example of this is Da Silva Neves et al. [47], in which English translations of the postulates of defeasible reasoning, proposed by Kraus, Lehmann and Magidor (KLM) [42], were tested in a survey. The translations were evaluated for plausibility on a linear scale by quantitatively judging participants' beliefs of the premises and the conclusion of each postulate. However, the methodology to measure the significance of the association between each component was vaguely defined. Their approach was thus irreproducible. In subsequent work [9], the authors used a refined methodology that combined theoretical and empirical analysis. Their material was the postulates of default reasoning, which dealt with rules like "generally, from it is sunny deduce that it is hot". They consider possibility distributions satisfying parameters in which the situation that both "it is sunny" and "it is hot" is true is more plausible than where "it is sunny" is true and not "it is hot" is true. In their empirical investigation, they checked three desirable properties for default reasoning against the KLM postulates for defeasible reasoning. Ragni et al. [52] have also taken a combined approach to investigate the link between formal theories of non-monotonic reasoning and the ability of humans to reason defeasibly. Their work compared the reasoning power of five classes of non-monotonic systems. In their theoretical analysis, they compared the predictions of each system on the Suppression Task [18], a logical experiment used in the psychology com-

munity in which subjects appear to retract valid logical inferences when the subjects gain new information. Additionally, they compared the predictions of each system when using strict knowledge and defeasible knowledge. Strict knowledge can be seen as monotonic in which the same conclusion always follows from the same premises whereas defeasible knowledge can be seen as non-monotonic in which the same conclusion typically follows from the same premises, but exceptions can occur. In their empirical analysis, Ragni et al. used three experiments that tested the predictions of each system as well as the inferences of human reasoners with strict and defeasible knowledge. The predictions were evaluated by multinomial process tree structures [7]. In previous work [5], we surveyed the postulates of three forms of defeasible reasoning using a chiefly empirical approach. The task for participants was to rank, with motivation, natural language translations of the postulates. Participants expressed their belief of each postulate on a graded Likert [23, 64] scale with positive and negative extremes. Our analysis used hit rates, a quantitative measure of the average participant response per postulate. The hit rates were a rudimentary measure of correspondence using a majority-rule approach. In the analysis of participant explanations, we found both normative and descriptive responses. While many studies investigate non-monotonic and defeasible reasoning with human reasoning, the relationship between belief change and human reasoning is less substantiated. Empirical work in this field is ongoing and evolving since no model of reasoning aptly reproduces human reasoning.

1.2 Research Hypothesis

Our first hypothesis is that human reasoning is consistent with postulates, advanced by Alchourrón, Gärdenfors and Makinson (AGM) [1], for belief revision. Our next hypothesis is that human reasoning is consistent with postulates, advanced by Katsuno and Mendelzon (KM) [41], for reasoning with belief update. We investigate the hypotheses at the postulate level and at the system level. We use the language of propositional logic in the formulation of our postulates. This is a language in which statements about the world are single propositions, e.g., "Anne is wearing a red dress" and "Duke is mowing the lawn". Each statement can be assigned a binary value, true or false More complex sentences of propositional logic are constructed using Boolean logic operators that act on or between propositions. Propositional

logic can also be applied to the general case where we consider formulas instead of concrete facts. Further, we note that once our hypotheses are tested using propositional logic as the underlying language, our results can be lifted to other forms of logic.

1.3 Dissertation Outline

We have published 2 papers on this work, during the course of writing this Master's thesis. The first publication, "Cognitive Defeasible Reasoning: the Extent to which Forms of Defeasible Reasoning Correspond with Human Reasoning" [5], was based on preliminary work and appeared in the conference proceedings of the Southern African Conference on Artificial Intelligence Research (SACAIR2020). The second publication, "Belief Change in Human Reasoning: An Empirical Investigation on MTurk" [6], will appear in the conference proceedings of SACAIR2021. This publication was a direct result of mature research from this thesis.

The outline of the rest of this dissertation is as follows. In Chapter 2 we present background material on the topics of propositional logic, possibility theory and belief change. In Chapter 3 we describe types of empirical research, statistical methods and ethical considerations. We then give an account of our experimental setup. In Chapter 4 we report on our experimental setup concerning our participants, experiment material, design, procedure, predictions and results. In Chapter 5 we analyse and discuss the results from our experiments and how they answer our research hypothesis. In Chapter 6 we conclude with a summary of our findings and suggest avenues for future work.

Chapter 2

Background

In this chapter, we discuss background material on propositional logic, belief change and the use of possibility theory as a framework for data validation.

2.1 Propositional Logic

Propositional logic is a formal language used in the AI community to represent knowledge and draw inferences from it. It deals with structures called propositions. We also refer to propositions as statements in this work. Examples of propositions are "Kyle drives a blue car", "It is sunny" and "The ice-cream shop is closed". The propositions in our examples are called atoms since they contain no inherent structure and cannot be decomposed into smaller propositions without giving up the original meaning. An important property of propositions is that each proposition is assigned a value of either true or false. We describe the essential components of the language of propositional logic next, with further details in Ben-Ari [8] and truth-functional properties of propositions discussed in Bergmann et al. [15].

A set of atoms, $\mathcal{P} = \{\mathbf{p}, \mathbf{q}, ..\}$, abbreviates a set of facts about a system. We can build more complex propositions, called formulas, from propositional atoms, using Boolean unary (negation) and binary (conjunction, disjunction, material conditional and material biconditional) operators. These operators are formalised in logic, but correspond to the natural language names as in the following: 'not' for negation, 'and' for conjunction, 'or' for disjunction, 'implies' for the material conditional and 'if and only if' for the material biconditional. The function of the negation operator is to give the value false

if and only if applied to a single formula whose value is true. Conversely, the negation operator gives the value true if and only if applied to a single formula whose value is false. The conjunction operator gives the value true if and only if applied to two formulas whose values are both true. The disjunction operator gives the value true if and only if applied to two formulas in which at least one of the formulas is true. The material conditional operator gives the value true if and only if applied to two formulas, in three cases. The first case is where the formulas' values are both true. The next case is where the formulas' values are both false. The last case is where the first part, the formula before the operator, is false and the second part, the formula after the operator, is true. The material conditional operator gives the value false otherwise. The material biconditional operator gives the true if and only if applied to two formulas whose truth values are the same, either both true or both false. The material biconditional operator gives the value false otherwise.

The propositional language $\mathcal{L}$ generated by $\mathcal{P}$ is the set of formulas defined recursively as follows:

- If $\alpha \in \mathcal{P}$ then α is a formula of $\mathcal{L}$;
- If α and β are formulas of $\mathcal{L}$, then so are $(\neg\alpha)$, $(\alpha \wedge \beta)$, $(\alpha \vee \beta)$, $(\alpha \rightarrow \beta)$ and $(\alpha \leftrightarrow \beta)$.

The precedence associated with these operators in descending order is: $\neg$, $\wedge$, $\vee$, $\rightarrow$ and $\leftrightarrow$. Parentheses may be used for disambiguation. As an example, consider a system described by three sentences. The first is "six divided by three is two", denoted by α. The second is "seven is an integer", denoted by β. The third is "eight is smaller than nine", denoted by γ. If the sentences α, β and γ are taken to be true, then by definition, the following are examples of formulas that are also true:

- $\neg\neg\alpha$
- $\alpha \wedge \beta \rightarrow \gamma$
- $\neg\alpha \rightarrow \beta \vee \beta \leftrightarrow \gamma$

Similarly, by definition, the following are examples of formulas that are false in our system:

- $\neg\alpha \wedge \gamma$

- $\neg(\alpha \vee \beta)$
- $\alpha \vee \neg\beta \leftrightarrow \alpha \rightarrow \beta \rightarrow \neg\gamma$

These examples are but a few that can be generated in our propositional language. Infinitely many formulas can be generated in our language, restricted only by time and computing resources. Now that we have defined the rules for the correct syntax of propositional formulas, we can assign meaning to it. In earlier examples using α, β and γ, we have shown syntactically correct formulas that could be generated in our language. Some of the formulas had the truth-value true and some had the truth-value false. The truth-value of a formula was determined by the Boolean operators and, importantly, the value of its constituents. In our examples, the truth-value of each constituent was given to be true. However, this is not always the case. We recall that each atom in a propositional language can have two truth values. The size of a propositional language is thus given by 2^n where n is the number of atoms. A formula in a propositional language can thus take on different truth values, depending on the values of its constituents. Furthermore, by assigning a truth-value to each atom in our language, we give it a meaning and obtain a picture of the state of our system. For every unique way of assigning truth values to all atoms of our language, we obtain another state in our system. This is referred to as an interpretation. Formally:

Definition 2.1 *(interpretation) An interpretation I is a total function that assigns one of the truth-values T or F to every atom in $\mathcal{P}$. Formally, I is given by $I : \mathcal{P} \longrightarrow \{T, F\}$. The set of all interpretations, also referred to as states or possible worlds, is denoted by W.*

Using interpretations, it is possible to check for combinations of constituents that make a formula true. Formally:

Definition 2.2 *(satisfiability) An interpretation I satisfies a formula α, written $I \Vdash \alpha$, if and only if one the following conditions hold:*

- $\alpha \in \mathcal{P}$ *and* $I(\alpha) = T$
- $\alpha = \neg\beta$ *and* I *fails to satisfy* β
- $\alpha = \beta \wedge \gamma$ *and* I *satisfies both* β *and* γ
- $\alpha = \beta \vee \gamma$ *and* I *satisfies at least one of* β *and* γ

- $\alpha = \beta \rightarrow \gamma$ *and* I *satisfies at least one of* $\neg\beta$ *and* γ
- $\alpha = \beta \leftrightarrow \gamma$ *and* I *satisfies either both* β *and* γ*, or else neither*

Model theory is used to study the relationship between theories and the structures in which the theories hold. We apply it to propositional logic to understand the relationship between propositional formulas and the structures in which it holds. We call these structures models. Formally, the set of interpretations that satisfy a formula α, written $\text{Mod}(\alpha)$, is called the set of models of α. α is satisfiable if and only if it has a model, i.e. iff $\text{Mod}(\alpha) \neq \emptyset$. α is valid if and only if $\text{Mod}(\alpha) = W$. α is equivalent to β, written $\alpha \equiv \beta$, if and only if both α and β have the same models, i.e. iff $\text{Mod}(\alpha) = \text{Mod}(\beta)$. α is unsatisfiable, or a contradiction, if and only if $\neg\alpha$ is valid if and only if $\text{Mod}(\alpha) = \emptyset$. Furthermore, we need a mechanism to compute inferences from a set of propositional formulas to enable reasoning. This can be done by comparing the models of two formulas, one of which is designated as the knowledge base and the other is designated as the inference. This mechanism is called entailment. A necessary condition for entailment is that both the knowledge base and inference must be known or contain the empty set. Formally,

Definition 2.3 *(entailment) A formula* α *entails another formula* β *or, expressed otherwise,* β *is a logical consequence of* α*, written* $\alpha \models \beta$*, if and only if* $Mod(\alpha) \subseteq Mod(\beta)$*. Similarly, a set of statements* Γ *entails another statement* β*, if and only if* $Mod(\Gamma) \subseteq Mod(\beta)$*.*

2.2 Belief Change: Revision and Update

We consider belief change as an umbrella term for both belief revision and update. These are operations defined for modifying beliefs when the world in which a reasoner exists experiences a static or dynamic change, respectively. We begin our review of belief change with an example, adapted from Fermé and Hansson [30]. We then discuss revision and update in detail as it pertains to our investigation.

Consider a set of sentences in natural language: "Cindy is a sportswriter at a media company" (α), "Carl is an accountant at a media company" (β) and "Two people who work in the same department of a company are co-workers" (γ). We assume this set represents all currently available information about

Cindy and Carl. Suppose that we receive the following piece of new information: "Cindy and Carl are co-workers" (δ). If we add the new information to our beliefs, then we obtain a new set of beliefs that contains the sentences α, β, γ and δ. We can define an operation of addition as one that takes a sentence and a set of previous beliefs and returns the minimal set that includes both the previous beliefs and the new sentence, i.e. $K = \{\alpha, \beta, \gamma, \delta\}$. This operation exemplifies the simplest way of changing a set of sentences. There are other more complex types of change. Suppose that upon visiting the media company, we learn that sportswriters work in the news department (ϵ) and accountants work in the finance department (ϕ). If we add ϵ and ϕ to K, the result will be a set with contradictory information: γ contradicts both ϵ and ϕ since Cindy and Carl are co-workers but work in different departments of the media company. The addition does not necessarily reflect the notion of a consistent revision. If we wish to retain consistency, then some subset of the original set must be discarded or a part of the new information should be rejected. In our example, there are several possible alternatives. The information about Cindy and Carl's work could be wrong, as could be the new information. It could be that Cindy and Carl work on the same project for part of the time or share a common department function. Finally, the claim that Carl and Cindy are co-workers could be wrong. Any of these options, either individually or combined, would allow us to solve the problem of incompatibility between the original and the new information or beliefs. Consequently, we can specify an operation that takes a set and a sentence and returns a new consistent set. The new set includes part or all of the beliefs in the original set and it also includes the new sentence, should we accept it. The outcome of a revision can be expressed as a consistent subset of the outcome of the addition.

AGM Postulates

Alchourrón et al. [1] produced influential work in their study of theory contraction and revision. Contraction is the process of reducing a set of sentences to take out a proposition while revision incorporates a proposition into a set of sentences. They investigated partial meet contraction functions and defined the basic postulates of these functions. They have shown that the properties of partial meet contraction functions, viz. closure, success, inclusion, vacuity, recovery and extensionality, satisfy the Gärdenfors rationality postulates [33] and that they are sufficiently general to provide a represen-

tation theorem for those postulates. An important outcome of their work is the properties and representation theorem for contraction functions, which has been extended to revision functions in later work. Katsuno and Mendelzon [40] model-theoretically analysed the semantics of revising knowledge bases by sets of propositional sentences. They provide a characterisation of all revision schemes that satisfy the Gärdenfors rationality postulates, in terms of minimal change for an ordering among interpretations.

We recall that Belief revision is a process of adapting one's beliefs with the motive that the world has changed statically. This means that a reasoner's previously valid beliefs should not be discounted. Rather, it indicates that new information has brought to light gaps in the reasoner's beliefs. We represent a reasoner's beliefs by a belief set, K, a binary revision operator, $*$, and a set of postulates for reasoning. K is a set of logically closed propositional formulas e.g. α, β, γ etc. We denote logical closure by the function C_n. The purpose of a revision operator $*$ is to take a reasoner's beliefs and a new piece of information, and return a set of beliefs that represents the result of the revision. There are eight postulates for reasoning in the AGM belief revision framework, shown in Table 2.1. Postulate R1 can be inter-

Table 2.1: AGM Postulates

R1.	$K = C_n(K)$ and $K * \alpha = C_n(K * \alpha)$
R2.	If $K * \alpha \models \beta$ then $K + \alpha \models \beta$
R3.	If $K \not\models \neg\alpha$ then (if $K + \alpha \models \beta$ then $K * \alpha \models \beta$)
R4.	$\alpha \in K * \alpha$
R5.	If $\alpha \equiv \beta$ then $K * \alpha \models \gamma$ iff $K * \beta \models \gamma$
R6.	If $\alpha \not\models \bot$ then $K * \alpha \not\models \bot$
R7.	If $K * (\alpha \wedge \beta) \models \gamma$ then $(K * \alpha) + \beta \models \gamma$
R8.	If $K * \alpha \not\models \neg\beta$ then (if $(K * \alpha) + \beta \models \gamma$ then $K * (\alpha \wedge \beta) \models \gamma$)

preted in two parts. The first says that a belief set K includes in it all its beliefs and the logical consequences of its beliefs. The second part says that when K is revised with new information α, the result is logically closed too. Postulate R2 says that the process of revising K with α is subsumed in the process of expanding K with α. The expansion of K with formula α, $K + \alpha$, is defined as the deductive closure of $K \cup \{\alpha\}$. Postulate R3 says that if it holds that the belief set is consistent with a formula α, then revising K by α is the same as expanding K with α. Postulate R4 says that α is con-

tained in the revision of K with α. Postulate R5 says that if α and β are logically equivalent formulas, then the revision of K with either yields the same result. Postulate R6 states that the revision of K with α is inconsistent only if α is inconsistent. Alchourrón et al. [1] submit that under transitive relations, their partial meet contraction and revision functions also satisfy Gärdenfors' [32] supplementary postulates, given by R7 and R8. Postulate R7 says that the revision of K with a conjunction $\alpha \wedge \beta$ is subsumed in the revision of K with the first conjunct and expansion with the second conjunct. Postulate R8 says that if that the revision of K with α is consistent with β, then the revision of K with the first conjunct and expansion with the second conjunct is subsumed in the revision of K with the conjunction $\alpha \wedge \beta$.

KM Postulates

Katsuno and Mendelzon [39] gave a characterisation of all revision methods that satisfy the AGM postulates in terms of a pre-order among models. In subsequent work [41], they defined postulates for updating a finite propositional knowledge base by partial orders or partial pre-orders over interpretations. The class of operators defined generalises Winslett's Possible Models Approach (PMA) [65], which Katsuno and Mendelzon argue is an update operator given that the PMA changes each world independently. Herzig and Rifi [37] used the KM postulates as a standard for evaluating ten different propositional update operations in the literature. In later work, Herzig et al. [36] studied a family of belief update operators by analysing the interplay between formulas and literals. They defined the operation of update as follows: first omit from the belief base every literal on which the input formula has a negative impact and then conjoin the resulting base with the input formula. They evaluated the update operators in two dimensions: the logical dimension, by checking the status of KM postulates, and the computational dimension, by identifying the complexity of several decision problems. The KM postulates have also been used by Miller and Muise [45] to evaluate a belief update mechanism for Proper Epistemic Knowledge Bases. This mechanism guarantees consistency in the knowledge base when new beliefs are added. More recently, Creignou et al. [21] argued that belief update within fragments of classical logic has not been addressed thus far. They investigated the behaviour of refined update operators concerning the satisfaction of the KM postulates and, in this context, highlighted the differences between revision and update. Ribeiro et al. [57] used the KM postulates to

define a class of belief update functions, called royal splinter functions, for non-finitary logics. In the following, we give an account of the KM framework as it pertains to our investigation of human belief change.

Belief update is a special case of belief change in which we record a change to our beliefs, rather than statically add information to it. As with revision, we represent a reasoner's beliefs by a belief set, K, a binary operator, $\diamond$ and a set of postulates for reasoning. As with revision, K refers to a logically closed set of propositional formulas. When we update K with new information μ, we accept that our representation of the world is insufficient. We must accommodate the new information into our beliefs, and not only the consistent parts. Update implies that our previous beliefs may have been flawed if the new information contradicts it. There are nine postulates in the KM belief update framework, shown in Table 2.2. We discuss them in the following. Postulates U1-U5 correspond directly to their counterparts for

Table 2.2: KM Postulates

U1.	$K \diamond \mu \models \mu$
U2.	If $K \models \mu$ then $K \diamond \mu$ iff K
U3.	If both K and μ is satisfiable then $K \diamond \mu$ is satisfiable
U4.	If K_1 iff K_2 and μ_1 iff μ_2 then $K_1 \diamond \mu_1$ iff $K_2 \diamond \mu_2$
U5.	$(K \diamond \mu) \wedge \phi \models K \diamond (\mu \wedge \phi)$
U6.	If $K \diamond \mu_1 \models \mu_2$ and $K \diamond \mu_2 \models \mu_1$ then $K \diamond \mu_1$ iff $K \diamond \mu_2$
U7.	If K is complete then $(K \diamond \mu_1) \wedge (K \diamond \mu_2) \models K \diamond (\mu_1 \vee \mu_2)$
U8.	$(K_1 \vee \mathsf{K}_2) \diamond \mu$ iff $(K_1 \diamond \mu) \vee (K_2 \diamond \mu)$
U9.	If K is complete and $(K \diamond \mu) \wedge \phi$ is satisfiable then $K \diamond (\mu \wedge \phi) \models (K \diamond \mu) \wedge \phi$

AGM belief revision. U1 says that a formula μ is a logical consequence of the result of updating a belief set with a formula μ. Postulate U2 says that if it holds that the belief set entails some formula μ, then updating the belief set with μ does not influence the belief set. A special case of Postulate U2 is when the belief set K is consistent. In this case, Postulate U2 is weaker than its counterpart, Postulate R2, in the AGM belief revision framework. Postulate R3 says that the result of an update is satisfiable if both the belief set and the formula μ, with which to update the belief set, is satisfiable. Postulate U4 allows for irrelevance of syntax when updating logically equivalent belief sets with logically equivalent formulas. Postulate U5 says that updating on μ

and conjoining the result with ϕ entails updating on the conjunction of μ and ϕ. Postulate U6 states that if updating on μ_1 entails μ_2 and if updating on μ_2 entails μ_1, then the updating with μ_1 has the same effect as updating with μ_2. The next postulate places a restriction on belief sets. Postulate U7 says that should some situation arise from updating a complete belief set with μ_1, and the same situation arises when updating with a different formula, say μ_2, then this situation must also arise from updating that belief set with the disjunction of the formulas, $\mu_1 \vee \mu_2$. Postulate U8 guarantees that every possible situation in the belief set is given independent consideration. The final postulate, U9, is again restricted to complete belief sets. It says that if updating a complete belief set with a formula μ and then conjoining the result with another formula ϕ is satisfiable, then updating with both formulas at once produces the same effect as updating with the first formula before conjoining the result with the second formula. In summary, Katsuno and Mendelzon [41] contrast revision and update as follows: a different ordering is induced by each model of the belief set for update while only one ordering is induced by the whole of the belief set for revision. The different orderings that result from an update reflect the different states of the world described by the belief set. For revision, the ordering corresponds to a single consistent state of the world.

2.3 Possibility Theory

Possibility theory has mathematical origins and is a theory capable of handling specific forms of uncertainty in incomplete information. It has been introduced by Zadeh [67], whose development of the theory was based on fuzzy sets. Possibility theory is an alternative to probability theory and differs from it by using necessity measures in addition to possibility measures. Possibility and necessity measures are represented as set-functions. A set-function [3, 26, 27] is a function whose domain is a family of subsets of some given set and that usually takes its values in the extended real number line, consisting of the real numbers $\mathbb{R}$ and $\pm\infty$. Possibility and necessity are measured on a scale between 0 and 1. For possibility, the scale represents the range from impossible to possible. For necessity, the scale represents the range from unnecessary to necessary. Dubois and Prade [25] give an extended overview of possibility theory and the derivation of its measures of possibility and necessity.

We present possibility theory as an approach for validating the responses in our experiments surveying the belief change postulates. In particular, we conducted experiments (refer to Experiment 3 and 4 in Section 3.2) in which postulates of belief revision and belief update were ranked by participants in terms of how plausible it was to them. The postulates were represented in a manner that enabled comparison with the interpretation of default rules in possibility theory. For context, we recall that Benferhat et al. [12] and Dubois and Prade [24] applied possibility theory to study postulates of non-monotonic reasoning. Benferhat et al. [11] showed that the possibilistic approach contributed a faithful representation of the KLM [42] postulates. Thus, a default rule "if a then b, generally", denoted $a \mid\sim b$, is represented by the constraint,

$$\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$$

on a possibility measure Π describing the semantics of the available knowledge. The constraint says that in the context where α is true, there exists situations where having β true is strictly more plausible than any situations where β is false in the same context. The details for the computation of the necessity and possibility measures for a set of default rules are given by Dubois and Prade [25]. In our approach, we formulated the AGM and KM postulates as material implication statements, or "if...then..." statements in English. Each postulate could then be decomposed into a set of premises (the part that follows "if") and a conclusion (the part that follows "then"). Participants in our experiments judged each component of the postulate separately. To enable comparison with the interpretation of defaults in possibility theory, we represented each material implication statement as a default rule, "generally from the premises, deduce the conclusion". In our computation of Π for each default rule, we used frequency counts obtained in the data analysis step of our experiments. The frequencies were reported in a 2×2 contingency table showing the number of positive and negative endorsements of the premises and the conclusion. The frequencies were tabulated in a contingency table format showing combinations of participant endorsements of the premises and the conclusion. For example, suppose the premises of AGM postulate R1 is the formula α and its conclusion is β. Then, the corresponding contingency table is indexed to find the combinations of α and β needed for Π. This means $\alpha \wedge \beta$ is the frequency of the endorsement of both the premises and the conclusion, while $\alpha \wedge \beta$ is the frequency of the endorsement of the premises and the the frequency of the non-endorsement of the

conclusion. A default rule was found plausible by English-speaking human reasoners in general if the condition $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ was met.

Chapter 3

Procedures and Methods

In this chapter, we give an account of different types of empirical research. We describe statistical methods used for data analysis in empirical research. We then outline our empirical approach. Lastly, we present ethical concerns and how we intend to manage them.

3.1 Types of Empirical Research

In the following, we discuss three traditional types of empirical research. The first is a quantitative design. Surveys and experiments are examples of quantitative research designs. The aim of surveys in this design is to provide a numeric description of the behaviour of a population, by looking at a sample of the population. On the other hand, experiments establish whether a particular treatment of its subjects influences an outcome. This treatment is typically applied to one group and witheld from another. The researcher then determines the performance of both groups with respect to the desired outcome. The next type of research design used for data collection is a qualitative design. Examples of qualitative designs are narrative research, phenomenology, grounded theory, ethnographies and case studies. Narrative research has origins in the humanities. In this type of research, a researcher makes observations about the lives of individuals and asks for stories about their lives [58]. The information is collated by the researcher who uses it to produce a chronological account of the events that took place in the individual's life. Phenomenology, a design of inquiry that stems from philosophy and psychology, is used by researchers to describe the lived experiences of

individuals about a shared phenomenon as told from the individuals' perspective. The aim is to generate a description that captures the crux of their experiences about the phenomenon. Another qualitative design is grounded theory, which has origins in sociology. With this design, the researcher uses the views of participants to theorise about a process, action or interaction. Ethnography is also an example of a qualitative design. It origins are in anthropology and sociology. The researcher studies the collective traits of a cultural group in their its natural habitat over time. Ethnographic data is collected through observations and interviews. The remaining type of research design is a mixed-methods design. This design is also known by different names in the literature, e.g., multimethod, quantitative and qualitative methods and synthesis, however, recent writings [2, 44, 46, 48] favour the term mixed-methods. This research design blends quantitative and qualitative designs in a study. Several mixed-methods designs exist. Convergent parallel mixed methods is a design in which the researcher brings together quantitative and qualitative data to generate a complete understanding of the research problem. The researcher typically collects both quantitative and qualitative data simultaneously. The data is then integrated into the interpretation of the overall results. As an alternative to integrating both types of data simultaneously, a researcher can use explanatory sequential mixed methods to conduct quantitative research first and then explain the results with qualitative research. The reverse sequence is employed in an exploratory sequential mixed-methods research. The qualitative phase of the exploratory design is used by the researcher to create an apparatus that fits the sample under observation, or to identify appropriate apparatuses and variables for the next quantitative phase. Mixed-methods research is preferred over quantitative or qualitative research alone, in five particular situations. The first is when it is used for the comparison of different perspectives from quantitative and qualitative data. The next is when explaining quantitative results with a qualitative follow-up data collection and analysis. Mixed-methods research is also beneficial for understanding experimental results through the perspective of the individuals in the study. A mixed-methods design develops a fuller picture of changes needed for underrepresented groups when qualitative and quantitative data are combined. Furthermore, it allows the researcher to have a better understanding of the needs and impact of an intervention program through collecting both types of data over time [22].

3.2 Our Approach

Our research hypothesis posits that human reasoning is consistent with the postulates defined for reasoning with belief change in terms of revision and update. To test this hypothesis, we conducted experiments in which human subjects are required to reason with natural language statements that resemble material implication rules, e.g. "If Jacob B is a truck driver then Jacob B does drive at night". Material implication rules have two components, a premise and a conclusion, and the association between each can be measured quantitatively and tested statistically. We approached our investigation through four experiments and give a description of the design, data collection, data analysis, data interpretation and data validation for each experiment in the following.

3.2.1 Experiment 1

Our first experiment aimed to generate a cogent knowledge base of English statements that could be used for reasoning.

Design

We followed a qualitative approach for this experiment. The task was to evaluate natural language material implication statements for clarity and bias.

Data Collection

Participants were recruited via a lottery on social media. Participants who consented to this experiment were asked to complete an evaluation spreadsheet and submit their completed evaluation to the researcher via email.

Data Analysis

The number of statements containing bias or poor language was tallied for each participant. Participants were required to justify their views.

Data Interpretation

We applied a baseline of 50% of participants as the minimum percentage for a statement to be rendered grammatically poor. We applied a baseline of 50% of participants as the minimum percentage for bias to be present in a statement.

Data Validation

Statements that exceeded a bias percentage of 50% were discarded and substituted with alternatives.

3.2.2 Experiment 2

Our next experiment aimed to divide the general knowledge base of English statements, obtained in the previous experiment, into two sets. The first set contained statements that were found plausible by our participants. The other set contained statements that were found implausible by our participants.

Design

We followed a mixed-methods approach for this experiment. The task was to rank the vetted statements from experiment 1 for plausibility on a linear scale from 1 (strongly disagree) to 5 (strongly agree) and provide an explanation for their answer. The explanation could be open-ended, where participants provide their own justification. Alternatively, participants could select an explanation from a range of options, provided to participants, corresponding to the interpretations of the material implication statement.

Data Collection

Participants have been recruited anonymously Amazon Mechanical Turk (MTurk, `https://tinyurl.com/ycks2v6a`). Participants who consented to this experiment were asked to complete a survey on Google Forms (`https://tinyurl.com/2p856vyh`).

Data Analysis

The rank for each statement was treated as quantitative data and the explanation was treated as qualitative data. We computed the average rank and the modal explanation for each statement.

Data Interpretation

We considered an average rank with a positive response (between 4 and 5) as criteria for plausibility. We reviewed explanations in terms of agreement with material implication interpretations. We aim to obtain a mixed set of plausible and implausible statements. This forms the general knowledge about the world that is assumed to hold for each participant in the subsequent experiments.

Data Validation

We discussed whether the modal explanation confirms or disconfirms the average rank for each statement.

3.2.3 Experiment 3

Our next experiment tested translations of the AGM postulates of belief revision for plausibility.

Design

We followed a quantitative approach for this experiment. The task was to rank concrete natural language rules corresponding to the AGM postulates of belief revision for plausibility on a linear scale from 1 (implausible) to 10 (extremely plausible). Each postulate was first decomposed into material implication rules to obtain a premise and conclusion. Participants had to rank the premises separately from the conclusion. No explanation was required.

Data Collection

Participants were recruited anonymously via MTurk. Participants who consented to this experiment were asked to complete a survey on Google Forms.

Data Analysis

The average rank for each rule corresponding to the premises and the conclusion, was computed. An average rank of 6 or greater indicated agreement with the rule. An average rank of 5 or less indicated disagreement with the rule. We computed a contingency table containing the frequencies of endorsements of the premises and the conclusion of each postulate. To test whether there is a relationship between the endorsement of the premises and endorsement of the conclusion, we conducted a hypothesis test. Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. It involves two hypotheses, null (H_0) and alternative (H_a). The null hypothesis specifies an assumption about a sample. The alternative specifies the opposite of the null. The test uses sample data to determine whether the null hypothesis should be rejected. Allowance for error must be made when conducting a hypothesis test because of the reliance on sample data to hypothesise about a population. A type I error occurs when a true null hypothesis is rejected. The other error, called a type II error, occurs when a false null hypothesis is accepted. In the determination of accepting a null hypothesis, measures for reducing error exist. An upper limit on the allowable probability of making a type I error can be specified. This is referred to as the significance level (α) of the test and is typically specified at 1% or 5%. To draw conclusions from a hypothesis test, a measure called the p-value can be used to determine how likely the sample results based on the assumption that the null hypothesis is true. The size of the p-value is directly proportional to the likelihood of the sample results: the smaller the p-value, the less likely the sample results. A comparison of the p-value is made with the significance level. The null hypothesis is rejected in favour of the alternate hypothesis if the p-value is less than α. The null hypothesis cannot be rejected otherwise. Hypothesis tests can also be used to determine if the correlation between variables in a study are statistically significant. For our experiment, we measured the association between the premises and conclusion using the phi-coefficient. The phi-coefficient is a linear measure of association between two variables: independent and dependent. The independent variable in our experiment is the endorsement of the premises. The dependent variable in our experiment is the endorsement of the conclusion. It has been introduced by Karl Pearson [20] and it also known as Matthew's Correlation Coefficient or the mean square contingency coeffi-

cient. To use a Phi-coefficient, a contingency table is required that shows the frequency counts of the independent and dependent variables. For example, consider a school of learners who analysed the following statement: "if ostriches are birds, then ostriches lay eggs". Their data is collated in Table 3.1.

Table 3.1: Contingency table of learner responses about the relationship between birds and laying eggs

	Ostriches lay eggs	Ostriches do not lay eggs	Total
Ostriches are birds	517	200	717
Ostriches are not birds	133	150	283
Total	650	350	1000

"Ostriches are birds" is a binary variable with values true and false. "Ostriches lay eggs" is also a binary variable with values true and false. We denote the top-left cell corresponding to a choice of "Ostriches are birds" and "Ostriches lay eggs" by the symbol A. We denote the top-right cell corresponding to a choice of "Ostriches are birds" and "Ostriches do not lay eggs" by the symbol B. The bottom-left cell corresponding to a choice of "Ostriches are not birds" and "Ostriches lay eggs" is denoted by the symbol C. The bottom-right cell corresponding to a choice of "Ostriches are not birds" and "Ostriches do not lay eggs" is denoted by the symbol D. The phi-coefficient formula for the 2×2 contingency table is thus:

$$\phi = \frac{AD - BC}{\sqrt{((A+B)(C+D)(A+C)(B+D))}} \tag{3.1}$$

The calculation of the phi-coefficient for our example gives $\phi = 0.237$.

Data Interpretation

The phi-coefficient produces a value between 0 and 1, extremes which denote no association and perfect association respectively. The closer the score to 0, the weaker the association between the premises and conclusion. The closer the score to 1, the stronger the association between the premises and conclusion. An error occurs when the denominator of the phi-coefficient evaluates to 0, because of any of the four sums in the denominator evaluating to 0. This is treated by setting the denominator to 1, an arbitrary value [61, 68].

The result is a phi-coefficient of 0 which is means that no association could be determined. Once computed, the phi-coefficient was tested for statistical significance using three tests: chi-squared, chi-squared with Yates' correction and Fisher's exact test. We set the significance level to 5% for each test. While there are arguments in favour of reducing the significance level to 0.05% [13, 43], 5% is typical in experimental psychology e.g. Hentschke and Stüttgen [35], Rana and Singhal [55] and Johnson et al. [38].

A chi-squared(or $\mathcal{X}^2$) test is a goodness-of-fit test used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table. We used a two-tailed or non-directional chi-squared test in our analysis since our comparison involves frequencies rather than a particular treatment effect. The formula for $\mathcal{X}^2$ is given by:

$$\mathcal{X}^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} \tag{3.2}$$

where O_i is the observed value and E_i is the expected value. In small contingency tables, or cells with values less than 5, the chi-square test alone is not reliable. In this case, Yates' correction for continuity [66] can be used to reduce the error of overestimating the statistical significance. The correction to the chi-squared test is obtained by subtracting 0.5 from the difference between each observed and expected value in a 2×2 contingency table. The formula for $\mathcal{X}^2_{Yates}$ is given by:

$$\mathcal{X}^2_{Yates} = \Sigma \frac{(|O_i - E_i| - 0.5)^2}{E_i} \tag{3.3}$$

where O_i is the observed value and E_i is the expected value. An alternative measure of goodness-of-fit is Fisher's exact test [31]. It determines a precise rather than approximate p-value, a measure of whether the difference in observed and expected frequencies are a result of random chance. It can be performed on any sample size, but is particularly useful for small sample sizes where approximations are inadequate. The formula for Fisher's exact test is given by:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{3.4}$$

where p is the p-value, the values of a, b, c and d are the values in a 2×2 contingency table and n is the total number of samples. It must be noted

that while some tests perform well on small sample sizes, a large sample warrants more confidence in the results. Thus, a comparison of the all three tests together provide us with a stronger position to assert the significance of the phi-coefficient than any of the tests alone.

Data Validation

As the postulates for belief revision intend to capture belief change in human reasoning, we assume that participants' reasoning will be consistent with it. However, we expect to find diverse responses and aim to discuss trends and inconsistencies in our sample. Our aim is to generalise whether the postulates for belief revision are consistent with English-speaking human reasoners in general. We compared each participants' response to the average response for the rule and we discussed the influence of age and gender on the rank of each rule. Further, we used the framework of possibility theory to validate whether the association for each postulate is significant in general, or whether the significance is specific to the concrete instances.

3.2.4 Experiment 4

Our final experiment tested translations of the KM postulates of belief update for plausibility. The design, data collection, data analysis, data interpretation and data validation are analogous to the previous experiment.

3.3 Ethics

It is important to observe that studies involving human subjects require significant preparation and the consideration of psychological and logistical challenges. Currently, in the context of the Covid-19 pandemic (World Health Organisation, `https://tinyurl.com/dw84a2v7`), new methods of safely and responsibly conducting empirical research must be observed, in compliance with ethical, legal and professional standards. We obtained ethical clearance from the Faculty of Science Ethics Research Committee at the University of Cape Town. We include the consent forms and a link to our data management plan in the Appendix.

3.3.1 Crowdsourcing and Mechanical Turk

Crowdsourcing is a means of gathering a response to a task from a large audience, usually via the internet. Google Forms, Qualtrics and SurveyMonkey are examples of tools for creating online surveys with a variety of question types. The surveys are typically distributed by email invitation or by sharing a unique URL to the survey. In terms of data collection for these tools, the onus is on the survey creator to identify participants and elicit responses. MTurk builds on the crowdsourcing concept by facilitating task creation, hosting tasks and managing data collection through its database of registered participants. Bentley et al. [14] found that crowdsourced surveys are completed faster and cost less than traditional surveys, with no significant ($< 10\%$) loss in accuracy. To ensure that responses are of acceptable quality, Grootswagers [34] suggests the inclusion of a trial task before the primary task to give participants an idea of what is expected from them. For the bulk of our reasoning experiments, we use Google Forms to design our surveys, and we use MTurk to crowdsource our data collection. MTurk uses unique terminology for its ecosystem of tasks, task providers and participants. A task is created by a task provider and is completed by participants. The task is called a Human Intelligence Task (HIT). A participant is referred to as a Worker. A task provider is referred to as a Requester. The identity of both the participants and task provider are hidden from one another. MTurk maintains a database of its Requesters and Workers, however. A Requester may make his contact information visible via the consent form of a HIT. A Requester that requires unique responses may ask for some form of identification from participants during the completion of a HIT. Workers complete tasks for monetary remuneration, called a reward, provided by the Requester. MTurk makes their service available at a base cost of 20% of the reward to participants. The minimum reward per task is \$0.01. MTurk allows the Requester to load reward credit in bulk or use Amazon's Billing Service to settle the outstanding fees once HITs are completed. While a survey offers a convenient format for eliciting participant responses, it is not the only type of HIT. HITs are a spectrum of tasks that require human input, internet access and a computer. Examples of other HITs include spell-checking, essay-writing and audio transcription. Tasks can also be created on external platforms, such as Google Forms, and a HIT can be created on MTurk with a URL to that task. There are several steps involved in the creation of a HIT before it can be posted. MTurk allows Requesters to choose participants according

to criteria such as location, age and number of previous HITs completed. In turn, when Workers access MTurk, they can search a list of available HITs and filter tasks by criteria such as the reward amount and completion time. MTurk gives Workers the option to preview a HIT and read its description before accepting it. An issue with anonymous data collection is that unsupervised participants tend to be less attentive than supervised participants. This is mitigated by checking in on participants at various stages of the HIT. The check-in aims to remind inattentive participants to pay more attention, typically through asking a trial question or manipulating an instruction [49]. When check-ins are conducted this way, there is less experimenter bias [50], subject crosstalk [28] and participant reactance. Springer et al. [62] recommend using a neutral non-persuasive tone in the language of the HIT to diminish the likelihood of selection bias based on persuasive language and target participant traits. HITs on MTurk are bound by a window of completion which can span a range of several minutes to several days. This means that participants are free to complete tasks when it is convenient for them, within the allotted time frame. The effect of participant performance on MTurk at different times of the day is robust, while there is some variation in participants' personality and prior experience across these recruitment times [4]. Another concern of mass anonymized data collection is accepting participant responses at face value. This is problematic because participants who depend on MTurk as a source of income are primarily incentivised by the monetary reward rather than interest in the HIT content. Participants also learn and apply behaviour from similar tasks which introduces unknown bias in their responses. MTurk has been shown to have limited ideological representation, in particular, where subjects hold more liberal attitudes than the general public [16]. The difference in convenience and probability sampling should be well-substantiated by repeat sampling over time rather than an analysis of background characteristics alone [19]. We used Mechanical Turk in previous work [5] to survey the correspondence of human reasoning with defeasible reasoning, belief revision and belief update respectively. We recorded a generally positive experience with a survey turnaround time below an hour for each participant and compensated Workers commensurate with the local minimum hourly wage.

3.3.2 Other Issues

Niche theory

The use of highly specialised theory and its application in an experimental setup places an additional responsibility on participants and influences the criteria for selecting participants. The main part of our investigation surveys English translations of logic-based postulates for reasoning with belief change. Our topic of study is to investigate empirically whether the postulates or are consistent with human reasoning. The main variable in our study is the plausibility, in part determined by our participants and in part determined by applying our data analysis procedure, of each postulate of belief revision and belief update respectively. We introduced hypothetical scenarios in natural language which corresponded to the structure of each postulate and asked participants whether they found the postulate plausible. We did not force our participants to adopt a particular position when asserting plausibility. Rather, we required that they provide their own view. Some postulates require the introduction of new information. In this case, we presented our participants with the background information and asked them how plausible they found the new information given the background information. By reasoning this way, participants have the same starting point. In other words, by making the knowledge about each postulate explicit, each participant had the same single source of knowledge from which to reason. This enabled us to infer how participants changed their beliefs to accommodate new information, given a common source of knowledge.

Handling of Participants

In experiments 2 to 4, we included a participant authentication question in the survey. This was done to verify the identity of our participants as human subjects. This was not necessary for the first experiment, as participants were recruited via social media platforms that linked a social media account to each participant. In experiments 3 and 4, we collected age and gender data. This allowed us to describe the population to which the belief revision and belief update findings generalise. Additionally, we created mock surveys and timed our responses to them. We predicted that each of the surveys would take a Worker between 30 and 45 minutes to complete. This time allocation also accounted for gaining access to the survey, reading the instructions and providing informed consent. Following best practice [34, 51, 62], we split

each of our surveys into multiple smaller surveys. Each smaller survey was uploaded as a HiT for Workers to complete. The shorter nature of the survey makes the content digestible and mitigates participant fatigue.

Handling of Mixed Data

There are several issues to consider when managing data in a mixed-methods design. In experiment 2, we used a mixed-methods approach to collect a rank corresponding to participants' beliefs about a general material implication statement and an explanation for it. In their explanation, participants indicated which component of the postulate they found plausible, the premises and/or the conclusion. Participants were also provided with a text field to supply additional comment. The quantitative and qualitative data were collected simultaneously from the same sample. In our treatment of the collected data, we computed frequency counts for the rank of each statement. We grouped our qualitative data into categories with common explanations. Our aim with the data interpretation was two-fold. First, we determined whether the explanations confirmed or contradicted the rank. Then, we determined whether the participants' explanations were consistent with interpretations of material implication statements in propositional logic. We discussed the influence of the quantitative data and qualitative data individually before concluding with our findings on plausible statements considering both types of data.

Chapter 4

Experiments

Our investigation involved four experiments with the task for each in the form of answering a survey.

4.1 Experiment 1

We prepared a survey of 30 general statements about the world for participants to evaluate for clarity and bias. The task was for participants to complete a table in which they identify statements whose language is unclear and statements which contain bias.

4.1.1 Participants

7 English-speaking participants were recruited for this experiment.

4.1.2 Material

Our participants were provided with instructions and a survey form. The instructions were given as in the following:

1. The purpose of this study is to evaluate general claims about the world for clarity (whether the claims make sense in English) and for bias (whether the claims are prejudiced or are not plausible).

2. The claims about the world made in this study take the form of "if...then..." statements e.g. "if Jacob B is a truck driver then Jacob B

does drive at night". It involves the behaviours and traits of people in various professions.

3. Your task is to read the statements numbered 1 - 30 in the Statement table and then answer questions (a) - (f) below in the Evaluation table provided.

In Table 4.1, we show the first part of the survey containing the questions for experiment 1. We show the material for experiment 1 in Appendix B,

Table 4.1: Evaluation table for experiment 1

No.	Question	Yes	No	Comment
(a)	I acknowledge that I am a native English speaker / that I have a good understanding and command of the English language.			
(b)	I acknowledge that I have read statements 1 - 30 in full.			
(c)	Do you believe the language used in any of the statements 1 - 30 to be unclear? If you answered "Yes", please write the statement number(s) in the "Comment" column and give a reason for each.			
(d)	Do you believe the content of any of the statements 1 - 30 to be biased or implausible? If you answered "Yes", please write the statement number(s) in the "Comment" column and give a reason for each.			
(e)	Are there any statements regarding behaviours or traits of people in a certain profession which you would like to see included in this survey? If you answered "Yes", please write your suggestions under the "Comment" column.			
(f)	Any feedback to add?			

from Figures B.1–B.3. In this experiment, reflexive statements of the form "if Noel W is a strong firefighter then Noel W is a strong firefighter", as seen in statements 6, 12, 18, 24 and 30, were included to determine whether

participants believe an "if...then..." statement when the premises is equivalent to the conclusion.

4.1.3 Design and Procedure

Participants were recruited via a lottery on social media on a first-come-first-served basis. Upon responding to the social media post advertising this survey, participants received an email. In the email, participants were provided with instructions, a consent form and a survey form. The survey form was a Microsoft Excel spreadsheet. Participants were allowed to complete the evaluation in their own time, with the requirement that they complete the evaluation within one week from receipt of the email. Each participant was remunerated with ZAR 50.

4.1.4 Predictions

We predicted that our participants would have sufficient general knowledge to understand the survey material. Given that the general statements about the world were estimated according to our perceived ideas about the world, we predicted that a moderate amount of statements would contain bias.

4.1.5 Experimental Results

After analysing our participants' responses, we found that the survey material was clear and written in unambiguous English. Our results also confirmed that bias was present in our material. In the following, we refer to two forms of bias: global statement bias and local statement bias. For global bias, the number of "Yes" responses to question (d) is used to calculate the global statement bias %. For local statement bias, the responses to question (d) given in the "Reason/Comment" column is used to calculate the local statement bias %.

Definition 4.1 *(global statement bias %) The global statement bias (GBS) % refers to the percentage of participants who responded "Yes" to question (d) of the survey. The formula for % global statement bias is given in Equation 4.1.*

$$\text{GBS } \% = \frac{\text{no. of "Yes" responses for question (d)}}{\text{total no. of participants}} \text{ X } 100 \qquad (4.1)$$

Definition 4.2 *(global statement bias threshold) The global statement bias threshold refers to a 50% global statement bias. For global statement bias to be present, the % global statement bias must be ≥ the global statement bias threshold.*

The GSB % is 85,71%. This means that 85,71% of participants expressed that at least one of statements 1 to 30 contained bias or was implausible. Since the GSB % exceeds the threshold of 50%, we concluded that global statement bias is present.

Definition 4.3 *(local statement bias %) The local statement bias (LSB) % for a statement S refers to the percentage of participants who both responded "Yes" to question (d) of the survey and specified a statement number S in the "Reason/Comment" column, where S = 1, 2, 3..., 30.*

$$\text{LSB } \% = \frac{\text{no. of "Yes" responses for question (d) for a statement S}}{\text{total no. of "Yes" responses for question (d)}} \text{ X } 100 \tag{4.2}$$

Definition 4.4 *(local statement bias threshold) The local statement bias threshold refers to a 50% local statement bias. For local statement bias to be present for a statement S, the LSB % must be ≥ the local statement bias threshold.*

The LSB% for statements 1, 3 and 23 is 66,67%. The LSB% for statements 2, 7, 8, 9, 15, 17, 22 and 25 is 50%. This means 11 statements contained bias or were found implausible. Since the LSB% for statements 1, 2, 3, 7, 8, 9, 15, 17, 22, 23, and 25 meets or exceeds the threshold of 50%, we concluded that local statement bias was present in these statements. Local statement bias was not present in the remaining statement. We found that local statement bias exists in 11 statements overall (36,67%). We deliberately did not prescribe the type of bias for participants to identify. Instead, we required that participants identify any form of bias that they perceive from the statements. One participant explained their choice of bias as the statement containing "structural bias". Another explained their choice of bias as the statement being a "presumption". Another explained their choice of bias as the statement being a "harmless bias". Another explained their choice of bias as the statement being a "caricature, popularised in the media". Another explained their choice of bias as the statement being a "stereotype". As the statements

in this survey inform the material for the following experiment, it was imperative to minimise the existing bias. As a result, we replaced the 11 biased statements with 11 new statements. These new statements were obtained by substituting statements containing local statement bias with a combination of statement suggestions from participants and our knowledge.

4.2 Experiment 2

In this experiment, we prepared a survey of 30 general statements about the world. The task was for participants to evaluate the degree to which they believe each of the statements in the survey and explain their answer.

4.2.1 Participants

30 English-speaking participants were recruited on MTurk for this experiment.

4.2.2 Material

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 explanations. In Appendix A in Figures B.4–B.6, we show the material. In these figures, statements in bold indicate that the original statement from Experiment 1 contained local statement bias and was replaced with a random combination of statement suggestions from participants and our knowledge. The survey material 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 survey URLs are provided in Appendix A.

4.2.3 Design and Procedure

Participants were recruited on MTurk and had to meet certain criteria. The first criterion was that a participant's *HIT Approval Rate (%) for all Requester's HITs* was ≥ 98. The next criterion was that the *Location* of all participants was restricted to the United States of America. The final criterion was that a participant's *Number of HITs Approved* was > 50. We show these settings in Figure 4.1 Personal characteristics such as age, gender, race

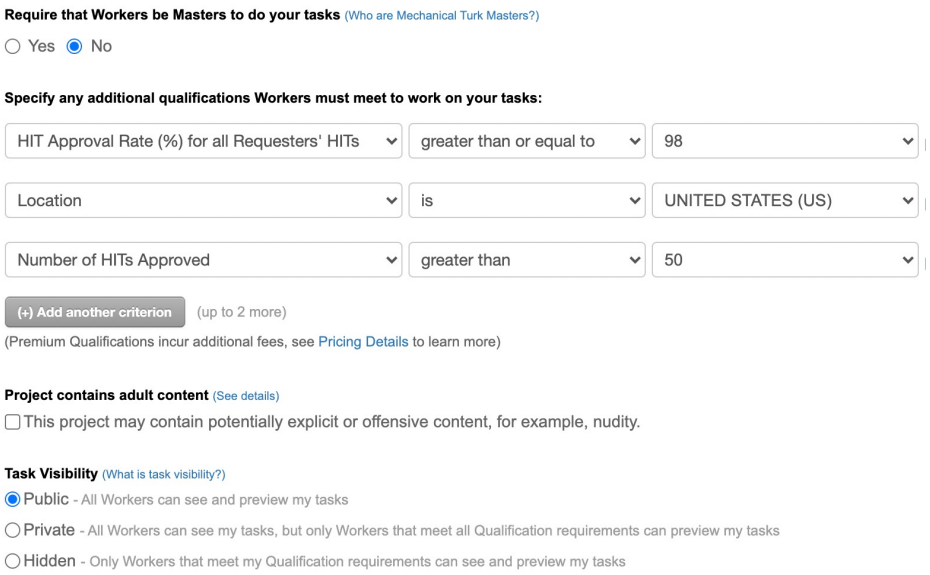

Figure 4.1: Worker requirements on MTurk for the general reasoning survey

and education level were not pertinent to this experiment and thus were not collected. Participants were not required to be Master workers. The *Task Visibility* requirement was set to *Public* such that all MTurk Workers could access our survey. The criteria in the remaining surveys for this experiment was identical to the first. The survey was published as a HIT on MTurk with 30 assignments available, corresponding to the number of required participants. Participants were not excluded from completing the other surveys in this experiment. In addition to the Worker requirements, our survey assignment on MTurk was published after following the steps outlined below. On the MTurk project landing site, we selected the *Survey Link* template and typed out the project details. The project details involved completing a

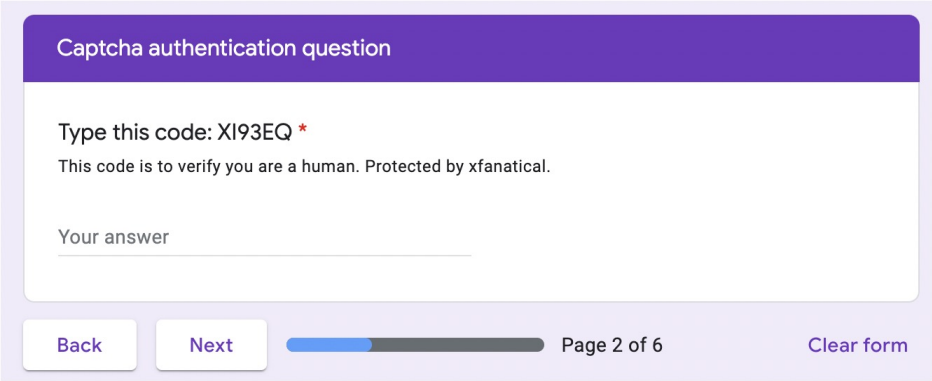

Figure 4.2: Participant authentication question for the general reasoning survey

form consisting of three tabbed panes: *Enter Properties*, *Design and Layout* and *Preview and Finish*. On the *Enter Properties* tabbed pane, there were three sections to complete: a general description of the survey to Workers, survey settings and worker requirements. For the description part, we included a title, a brief description of the survey content and key words. For the survey settings part, we specified the fields *Reward per response* as USD 2,5, the *Number of respondents* as 6, *Time allotted per Worker* as 1 hour, *Survey expires in* as 7 days and *Auto-approve and pay Workers in* as 3 days. We show a screenshot from MTurk with these settings in Figures 4.1. On the *Design and Layout* tabbed pane, we provided the URL to our survey, instructions for our participants and a text area for participants to submit their HIT completion code. On the *Preview and Finish* tabbed pane, MTurk presented a summary of the task as it appeared to Workers. We finished the creation of our project by clicking the *Finish* button on this pane. This process of creating a new project on MTurk was applied five times, one for each survey. In addition, we imposed a batch requirement of 6 responses at a time mitigating additional costs from MTurk. This meant that each survey was published 5 times to obtain 30 responses. We describe the participant procedures for the first of two belief revision surveys, next.

Once an assignment was accepted, participants were asked to complete a survey on Google Forms. The first page of the survey was the consent form. Participants were informed that they could expect to agree to the terms and conditions in the consent form before proceeding with the survey. Next, participants were informed that they could expect to pass a text-based

authentication question by repeating a randomly generated alpha-numeric code (Figure 4.2 and that they could expect to pass a simple English language proficiency question. In the main part of each survey, we provided 6 general "if...then..." statements about the world about the behaviours and traits of people in various professions. The main task was to indicate the degree to which they believed each statement on a scale from 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 explanations. At the end of the survey, we provided participants with a unique survey code that participants would need to enter on MTurk to receive payment. As a means of identifying individual responses, we asked participants to provide their unique MTurk identification number (the MTurk terminology is *WorkerID*). Each survey had a reward of USD 2,5 (at an exchange rate of 1 USD = 16 ZAR) and payment was approved within 3 days of completion. Although participants were given 1 hour to complete each survey, it was expected that they might complete it within 20 minutes. We gave no bonus reward for the surveys.

4.2.4 Predictions

We predicted that the strength of our participant's beliefs will differ across statements because participants have different backgrounds and because participants were not restricted by age, gender or other personal characteristics. We predicted that our participants' explanations would be consistent with the interpretations of the statements as a material implication in classical logic.

4.2.5 Experimental Results

We downloaded the survey responses from Google Forms and used Microsoft Excel for coding. Upon inspecting our data, all responses were found suitable for data analysis. In the following, we present our quantitative and qualitative analysis separately before discussing whether the statements in this experiment were found plausible by our participants.

In our quantitative analysis, we consider as quantitative data the rank that a participant has given for a statement. We computed the frequency of each belief (number of 1s, 2s,...,5s) for each statement. We used a chi-squared

test to investigate our null hypothesis: no relationship exists between the statement number and the respondents' beliefs. The chi-squared test statistic $\chi^2 = 348{,}922$ is greater than the critical value $\chi^2 = 142{,}138$. Therefore, the chi-squared test does fall in the rejection region. Thus, the null hypothesis is rejected in favour of the alternative hypothesis at a 5% significance level. This means that there is a difference in our respondents' beliefs across the 30 statements. This result is consistent with our prediction that 30 randomly selected participants will not reach the same conclusion for every statement. This result also means that this difference needs to be analysed further. In the next step of our analysis, we determined the mean rank for each statement using the formula: mean rank $= \frac{\text{sum of individual beliefs}}{\text{total number of participants}}$. The mean rank is a number representing the overall belief of a statement by our participants. We plot the average belief for each statement in Figure 4.3. The minimum rank

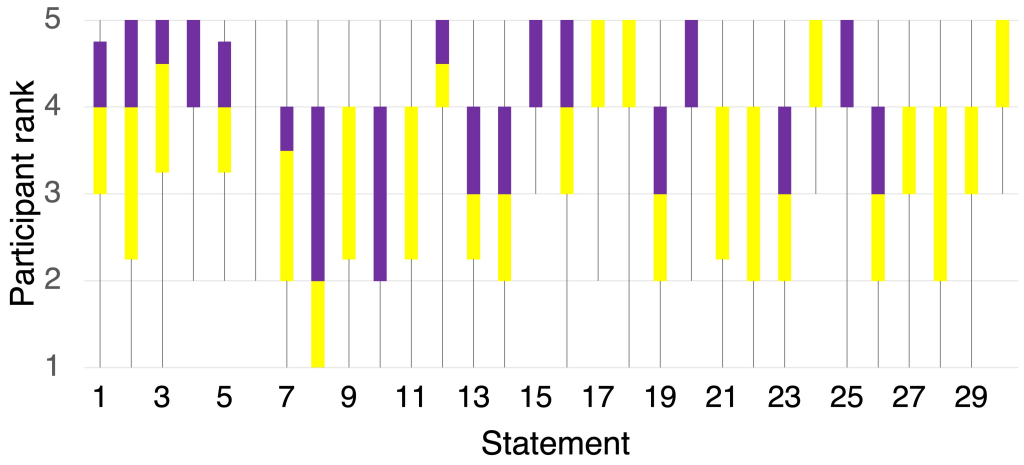

Figure 4.3: Box plot of statement rank

a statement can receive is 1 and the maximum rank is 5. The mean rank for all statements ranges from the lowest point of 2,37 (statement 8) to the highest point of 4,63 (statement 6). To explore the variability of respondents' beliefs with respect to the sample mean, we have calculated the coefficient of variation (CV) for each statement. Overall, the CV for all statements ranges from 13,67% (statement 30) to 63,13% (statement 8). Only one statement has significant variability of greater than or equal to 50% (statement 8, CV = 63,13%). Further, our goal is to find the statements that the majority of our participants perceive as plausible. A plausible statement is referred to

as a statement with a mean rank ≥ 4. According to this criteria, 30% (9 statements) were found plausible while 70% (21 statements) were not. The plausible statements in the survey are 4, 6, 12, 15, 17, 18, 24, 25 and 30. The majority of the plausible statements have the form "If A is the case then A is the case". This means that participants find a statement plausible given that the premises and conclusion are equivalent. The non-plausible statements in the survey are 1-3, 5, 7-11, 13-14, 16, 19-23 and 26-29. These statements involved different premises and conclusions. We argue that the subjective nature of these statements has divided the views of our participants, which is not unexpected. However, since the majority of statements in this survey were found non-plausible, we have had to modify our set of statements to obtain a more balanced set before using it in further experiments.

In our qualitative analysis, we consider as qualitative data the explanation that a participant has given for their rank of a statement. The explanations were represented in a mixed format comprising 5 options: a set of 4 pre-determined explanations and 1 open-ended explanation. The pre-determined explanations correspond to combinations of endorsement of the premises and conclusion of each statement. The open-ended explanation allowed participants to express their views, in addition to or in place of the pre-determined explanations. No restriction was placed on how participants selected explanations from the available options. As a result, the data we collected consists of 31 combinations of explanations, called explanation categories, for each statement. We computed the frequencies of explanation categories for each statement. In Figure 4.4, we show the relative frequency (RF) % of explanation categories across all significant (>0%) explanations. A, B, C and D correspond to individual explanations endorsed by participants. A refers to an endorsement of both the premises and conclusion of a statement. B refers to an endorsement of the premises, but not the conclusion. C refers to an endorsement of the conclusion, but not the premises. D refers to endorsing neither the premises nor the conclusion. As reported previously, participants were allowed to select multiple explanations. This is indicated in the graph as conjunctions of singular explanations e.g. A&B&D for the case of endorsing explanations A, B and D. Further, the category Other refers to an open-ended explanation provided by a participant. The explanation categories that were supported by our participants include, in descending order RF %: A, B, D, C, A&B&D, A&B, A&D, B&D, C&D, A&C, B&D, Other, A&B&C&D, A&B&Other, A&C&D, B&Other, A&B&C, A&Other and B&C&D. Overall, the 4 most frequently used explanations in order from first to fourth are A,

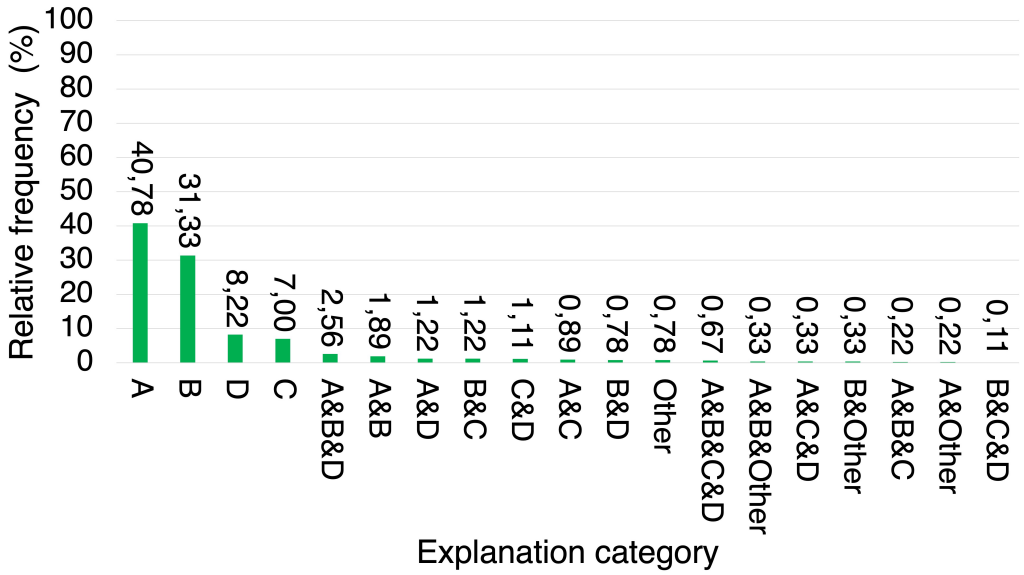

Figure 4.4: Bar plot of explanation category against relative frequency (%)

B, D and C. The majority of our participants (87,33%) preferred endorsing a single pre-determined explanation over endorsing multiple explanations or providing their own. We determined the modal explanation category, the explanation category with the highest frequency, for each statement. We observed a unique modal explanation category for each statement. 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 endorsement of the conclusion. A is the modal explanation category for statements 1-7, 9, 11-12, 15, 17-18,20,24-25 and 30, with 19, 14, 18, 12, 13, 22, 13, 15, 11, 17, 18, 18, 19, 14, 18, 17 and 20 responses respectively. B is the modal explanation category for statements 8, 10, 13-14, 16, 19, 21-23 and 26-29, with 17, 11, 13, 11, 14, 13, 15, 17, 14, 14, 11, 13 responses respectively.

In our combined analysis, we consider the plausibility of the statements in terms of both their mean rank and corresponding explanation. An explanation is defined as plausible if and only if the participant, in their explanation, endorses the premises but does not endorse the conclusion. In the case where a participant does endorse the premises but not the conclusion, we refer to it as a logical violation. This is consistent with the interpretation of mate-

rial implication statements ($\alpha \rightarrow \beta$ or if α then β) in classical propositional logic. As we have reported in our quantitative analysis, statements 4, 6, 12 15, 17, 18, 24, 25 and 30 are found plausible with a mean rank of ≥ 4. We analysed the corresponding explanations to determine whether or not the explanations confirmed or disconfirmed the mean rank. The frequency of logical violations committed by our participants ranges from 16,67% to 40% for the 9 plausible statements. This number is less than our baseline of 50%. Thus, the participants' explanations are consistent with our analysis of the mean rank for statements 4, 6, 12 15, 17, 18, 24, 25 and 30. As a result, the rank and explanations for these statements confirmed the statements' plausibility. We have also reported that the remaining 21 statements were found non-plausible in our quantitative analysis. Similarly, we analysed the corresponding explanations to determine whether or not the explanations confirmed or contradicted the mean rank. The frequency of logical violations committed by our participants for these statements ranges from 20% to 70%. In the cases of statements, 1-3, 5, 7, 9, 11, 13, 16, 20 and 28-29, the frequency of logical violations is less than our baseline of 50%. Thus, the participants' explanations for these statements are consistent with our analysis of their mean rank. While the explanations seemed to indicate plausibility, the rank did not. This contradiction needs to be assessed further and we provide a discussion of the open-ended responses for this experiment in our project repository, linked in Appendix C. Additionally, in the cases of statements 8, 10, 14, 19, 21-23 and 26-27, the frequency of logical violations is greater than or equal to our baseline of 50%. Thus, the participants' explanations for these statements were consistent with our analysis of their mean rank. Thus, the participants' explanations are consistent with our analysis of their mean rank for statements 8, 10, 14, 19, 21-23 and 26-27. As a result, the rank and explanations for these statements confirmed the statements' non-plausibility.

4.3 Experiment 3

In this experiment, we prepared a survey of English statements corresponding to translations of the AGM postulates for belief revision. The task was for participants to evaluate the degree to which they believe each statement in the survey.

4.3.1 Participants

50 English-speaking participants were recruited on MTurk for this experiment.

4.3.2 Material

The postulates of belief revision tested in this experiment are the 8 AGM postulates: *Closure (R1)*, *Inclusion (R2)*, *Vacuity (R3)*, *Success (R4)*, *Extensionality (R5)*, *Consistency (R6)*, *Super-expansion (R7)*, *Sub-expansion (R8)*. Each postulate was decomposed into its premises and conclusion, which we refer to as rules. The postulates were then reformulated as material implication rules of the form "if the premise holds then the conclusion holds". We show the decomposition and reformulation of each postulate in Tables 4.2–4.9. Altogether, there are 18 rules consisting of either premise rules or conclusion rules. We excluded rules that are tautologies from our survey and assumed that our participants would find tautologies plausible. The next step was to translate each rule into English. An equal, but random mix of plausible and non-plausible statements from Experiment 2 was used for this step. Once instantiated, the rules were included in the survey and the task was for each participant to rank the plausibility of the concrete rule on a linear scale from 1 (implausible) to 10 (extremely plausible). We divided the 18 rules between two surveys with an equal but smaller number of rules. We show the survey material for this experiment in the Appendix A in Figures B.7–B.10. The survey URLs are provided in Appendix A.

Table 4.2: Decomposition of Postulate R1

Postulate	$K * \alpha = C_n(K * \alpha)$
Material implication rule 1	If $\top$ then $K = C_n(K)$
Premise rule	$\top$
Conclusion rule	$K = C_n(K)$
Material implication rule 2	If $\top$ then $K * \alpha = C_n(K * \alpha)$
Premise rule	$\top$
Conclusion rule	$K = C_n(K)$

Table 4.3: Decomposition of Postulate R2

Postulate	If $K * \alpha \models \beta$ then $K + \alpha \models \beta$
Material implication rule	If $K * \alpha \models \beta$ then $K + \alpha \models \beta$
Premise rule	$K * \alpha \models \beta$
Conclusion rule	$K + \alpha \models \beta$

Table 4.4: Decomposition of Postulate R3

Postulate	If $K \not\models \neg\alpha$ then (if $K + \alpha \models \beta$ then $K * \alpha \models \beta$)
Material implication rule	If $K \not\models \neg\alpha$ then (if $K + \alpha \models \beta$ then $K * \alpha \models \beta$)
Premise rule	$K \not\models \neg\alpha$
Conclusion rule	If $K + \alpha \models \beta$ then $K * \alpha \models \beta$

4.3.3 Design and Procedure

Participants were recruited on MTurk and we established an internal set of criteria that they needed to satisfy. The criteria set on MTurk for the first survey, testing the concrete rules for AGM postulates R1 to R4, are shown in Figure 4.5. The first criterion was that a participant's *HIT Approval Rate (%) for all Requesterś HITs* was greater than or equal to 98. The next criterion was that a participant's *Number of HITs Approved* was greater than 50. Participants were not required to be Master workers. The *Task Visibility* requirement was set to *Hidden* to ensure that only participants who met our criteria on MTurk were able to view and participate in our tasks. Personal characteristics such as age and gender were collected in this survey. This requirement was recognised in the consent form of the survey and only requested once a participant had provided their consent to the survey on Google Forms. As a result, demographic data was not shared with MTurk. The criteria set on MTurk for the second survey, testing the concrete rules for AGM belief revision postulates R5 to R8, are shown in Figure 4.6. These criteria were identical to the first survey. However, we needed a manner in which to grant access only to those participants who had completed the first survey, to avoid duplicate responses and to ensure that Workers were tested on all 8 AGM postulates. To do this, we created two new qualification types on MTurk: *Completed_Part_1* and *Completed_Part_2*. Each qualification type also has an associated numeric score, between 0 and 100, that could be assigned to Workers. We assigned the qualification types *Completed_Part_1*

Table 4.5: Decomposition of Postulate R4

Postulate	$\alpha \in K * \alpha$
Material implication rule	If $\top$ then $\alpha \in K * \alpha$
Premise rule	$\top$
Conclusion rule	$\alpha \in K * \alpha$

Table 4.6: Decomposition of Postulate R5

Postulate	If $\alpha \equiv \beta$ then $K * \alpha \models \gamma$ iff $K * \beta \models \gamma$
Material implication rule	If $\alpha \equiv \beta$ then $K * \alpha \models \gamma$ iff $K * \beta \models \gamma$
Premise rule	$\alpha \equiv \beta$
Conclusion rule	$K * \alpha \models \gamma$ iff $K * \beta \models \gamma$

with score 100 and *Completed_Part_2* with score 0 to indicate that a specific Worker has completed the first survey, but has not completed the second survey. Upon completing the second survey, we updated a participant's *Completed_Part_2* score to 100. In addition to the Worker requirements, our survey assignment on MTurk was published after following the steps outlined below. We created a new project on MTurk for each survey. On the project landing site, we selected the *Survey Link* template and typed out the project details. The project details involved completing a form consisting of three tabbed panes: *Enter Properties*, *Design and Layout* and *Preview and Finish*. On the *Enter Properties* tabbed pane, there were three sections to complete: a general description of the survey to Workers, survey settings and worker requirements. For the description part, we included a title, a brief description of the survey content and key words. For the survey settings part, we specified the fields *Reward per response* as $3,75, the *Number of respondents* as 50, *Time allotted per Worker* as 1 hour, *Survey expires in* as 7 days and *Auto-approve and pay Workers in* as 3 days. We show screenshots from MTurk with these settings in Figures 4.5–4.6. On the *Design and Layout* tabbed pane, we provided the URL to our survey, instructions for our participants and a text area for participants to submit their HIT completion code. On the *Preview and Finish* tabbed pane, MTurk presented a summary of the task as it appeared to Workers. We finished the creation of our project by clicking the *Finish* button on this pane. This process of creating a new project on MTurk was applied twice, one for the survey testing AGM postulates R1 to R4 and one for the survey testing AGM postulates R5 to R8. We describe

Table 4.7: Decomposition of Postulate R6

Postulate	If $\alpha \not\models \bot$ then $K * \alpha \not\models \bot$
Material implication rule	If $\alpha \not\models \bot$ then $K * \alpha \not\models \bot$
Premise rule	$\alpha \not\models \bot$
Conclusion rule	$K * \alpha \not\models \bot$

Table 4.8: Decomposition of Postulate R7

Postulate	If $K * (\alpha \wedge \beta) \models \gamma$ then $(K * \alpha) + \beta \models \gamma$
Material implication rule	If $K * (\alpha \wedge \beta) \models \gamma$ then $(K * \alpha) + \beta \models \gamma$
Premise rule	$K * (\alpha \wedge \beta) \models \gamma$
Conclusion rule	$(K * \alpha) + \beta \models \gamma$

the participant procedures for the first of two belief revision surveys, next.

Upon accepting the HIT and gaining access to our first survey, participants were asked to authenticate themselves. This was done through an image-based trick question. A screenshot of this question is included in Figure 4.7. This question was introduced to prove that participants were human and not computers. Upon successfully identifying the cat amongst a set of images including three rabbits, participants were allowed to continue with the survey. Upon selecting the incorrect message, a feedback window was displayed and participants were allowed to redo the authentication question. By managing the authentication this way, we encouraged our participants to be alert and pay attention to their responses. The next part of the survey was the consent form which included an invitation to participate, information about procedures, recording, risks and feedback, a disclaimer, a confidentiality agreement and information about what signing the consent form means. Upon signing the consent form, participants were allowed to proceed with the survey. The next part of the survey asked participants to confirm that they satisfy the survey inclusion criteria, that they 1) are 18 years old or above, 2) reside in the United States of America and 3) are fluent in English. The next part of the survey asked participants to provide their age and gender. The next part of the survey listed a set of key words and their explanation, introducing the topic matter of the survey. Participants were advised to read the key words carefully before proceeding with the rest of the survey. The key words are included in the project repository, linked in Appendix C.

At any point in the survey, participants were allowed to navigate back

Table 4.9: Decomposition of Postulate R8

Postulate	If $K * \alpha \not\models \neg\beta$ then (if $(K * \alpha) + \beta \models \gamma$ then $K * (\alpha \wedge \beta) \models \gamma$)
Material implication rule	If $K * \alpha \not\models \neg\beta$ then if $(K * \alpha) + \beta \models \gamma$ then $K * (\alpha \wedge \beta) \models \gamma$
Premise rule	$K * \alpha \not\models \neg\beta$
Conclusion rule	if $(K * \alpha) + \beta \models \gamma$ then $K * (\alpha \wedge \beta) \models \gamma$

to the key words during the survey. The key words were also placed in a Google Docs document and linked as a URL in a text area above each rule, for ease of access to the key words. Next, participants were provided with a practice question. This was included to help participants understand the task expected of them. Passing the practice question required participants to select a value on a linear 10-point scale, from 1 = extremely implausible to 10 = plausible, corresponding to how plausible they find the given rule. For the practice question, participants were provided with a feedback prompt indicating success if they had selected a value on the scale from 1 to 10 corresponding to how plausible they found the rule in the example. Each rule in the survey was marked as *required (*)*. In the main part of the survey, participants were provided with the concrete rules for AGM postulates R1 to R4 and were asked to rate its plausibility, as in the practice question. The main part of the survey was estimated to take 10 to 20 minutes at a leisurely pace. At the end of the survey, participants were provided with a unique survey code that they needed to enter on MTurk to receive payment. Participants were informed that they needed to provide their MTurk WorkerID before submitting the survey. There was no bonus reward for this survey. Participants were also informed that this survey was the first part of a two-part survey. Participants were asked to indicate their interest on the last page of the survey by selecting "Yes" when prompted. At this point in the survey, participants were also given a survey batch number, enabling them access to the second survey, should they consent to it. Although participants were given 60 minutes to complete the first survey, it was expected that they might complete it within 20 to 40 minutes. At the end of the survey, participants were thanked for their time and were allowed to provide feedback about their experience in this survey. The second part of the survey followed the same procedure as the first part, except for a few differences.

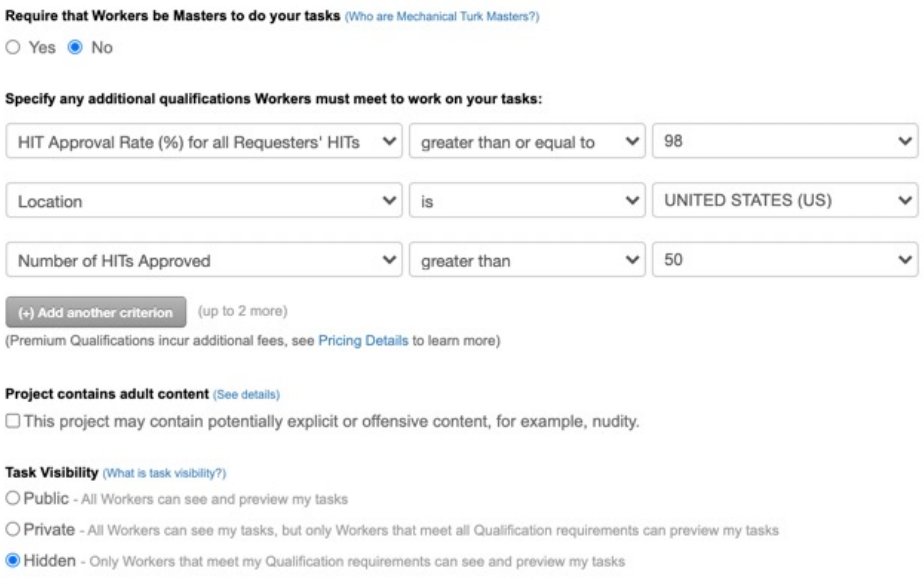

Figure 4.5: Worker requirements on MTurk for the AGM belief revision survey testing postulates R1 to R4

The authentication question asked participants to identify a bicycle in a set of images containing motorcycles. A screenshot of this question is included in Figure 4.8. No practice question was included in the second survey. The main part of the survey contained the concrete rules for AGM postulates R5 to R8.

4.3.4 Predictions

We aimed to determine whether the endorsement of the premises was preferentially associated with the endorsement of the conclusion of each postulate. We computed a contingency table that corresponds to the combinations of endorsement of the premises and conclusion in classical logic. To identify in which cell of the contingency table a participant's response matched, we applied the logical evaluation of the implication, premises $\longrightarrow$ conclusion. The degree of association between the endorsement of the premises and the endorsement of the conclusion was computed using the phi-coefficient (ϕ) and statistically tested using chi-squared ($\mathcal{X}^2$). For each postulate, the statistical hypothesis H_0 (null hypothesis) was "there is no significant association

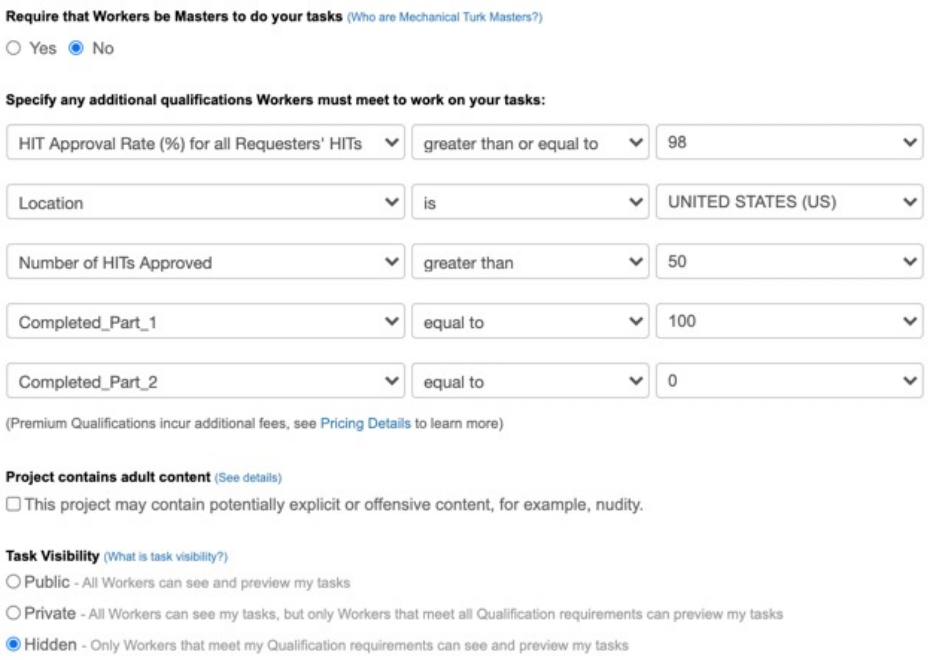

Figure 4.6: Worker requirements on MTurk for the AGM belief revision survey testing postulates R5 to R8

between the endorsement of the premises and the endorsement of the conclusion". The alternate hypothesis, H_1, was "there is a significant positive association between the endorsement of the premises and the endorsement of the conclusion". 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 [35, 38, 55]. Under the general hypothesis that the participants' inference tends to corroborate the 8 AGM postulates, we predicted that H_0 would be rejected for each postulate. Under the hypothesis that participant judgements are consistent with the AGM postulates, we predicted a high proportion of participants would not commit any logical violation.

4.3.5 Experimental Results

We downloaded the survey responses from Google Forms and used Microsoft Excel for coding. The rank given by each participant was coded as either True or False. The code True was assigned to statements with a plausi-

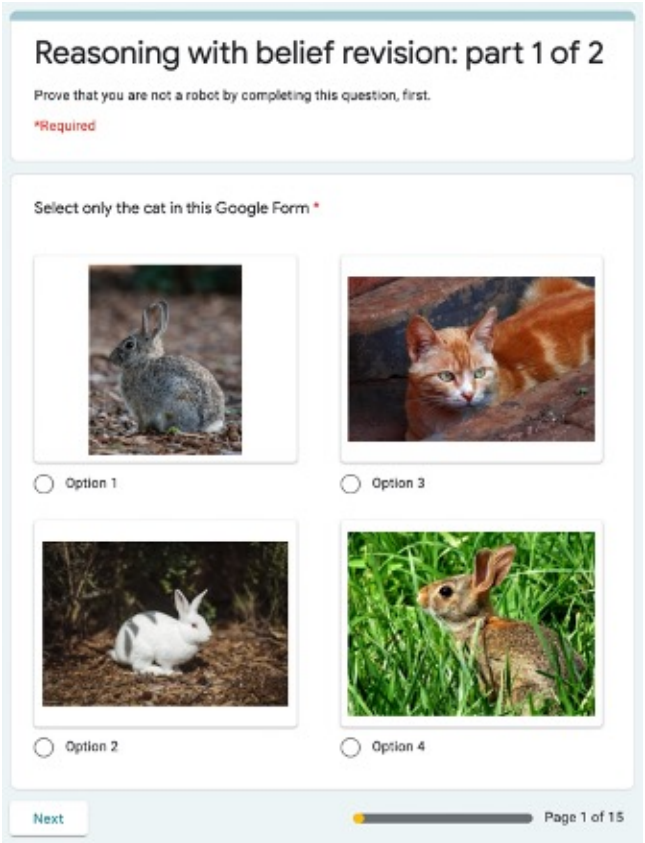

Figure 4.7: Participant authentication question for AGM belief revision survey testing postulates R1 to R4

bility rating of > 5 and ≤ 10. The code False was assigned to statements with a plausibility rating of ≥ 1 and < 6. Where premise or conclusion rules contain material implication or material equivalence statements, for example, R3, R5, and R8, these rules were decomposed further. The decomposition process was applied until the premise rule and conclusion rule for each postulate involved only a single material implication statement. Each rule obtained by this decomposition process was tested in the survey. In our coding process, rules were grouped according to the part of the postulate that it represented, the premises or the conclusion. The overall value for the premises of a postulate was obtained by the logical valuation of the conjunction of all the codes used in the rules for the premises. The same process was used to obtain the overall value for the conclusion. In the case of the premise

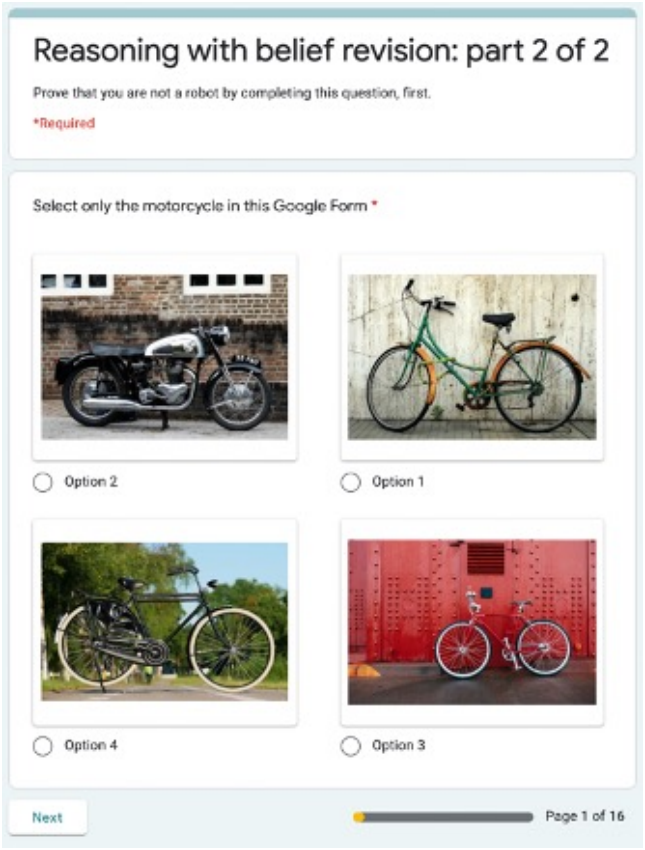

Figure 4.8: Participant authentication question for AGM belief revision survey testing postulates R5 to R8

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 the conclusion. During the data collection process, 50 participants completed the first belief revision survey. Although each participant expressed interest in taking the second belief revision survey, only 35 participants responded to the second survey. As a result, our adjusted sample size for each of the surveys is 35 participants. This was done to ensure that every participant evaluated each AGM postulate. Once the data was collected, a data cleaning process was performed. This step was programmed in R and the script is available in our project repository, linked in Appendix C. An inspection of the data from MTurk revealed that the *Lifetime Approval Rate for Requester's 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 the 35 participants were divided by gender as 22 male (62,86%) and 13 female (37,14%), as shown in Figure 4.9. Participants ages ranged from 22 to 74 years old. In Figure 4.10, the distribution of participant age by gender is shown 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. The box plot for female participants is higher than the box plot for male participants, suggesting that although there are fewer male participants than female participants overall, the female participants showed a greater spread in age than their male counterparts. The four sections of the box plot for male participants are uneven, as the four sections of the box plot for female participants. We show a box plot of the participant rank for our 18 concrete

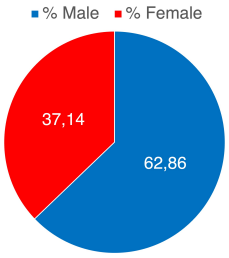

Figure 4.9: Pie chart of participant gender distribution by percentage

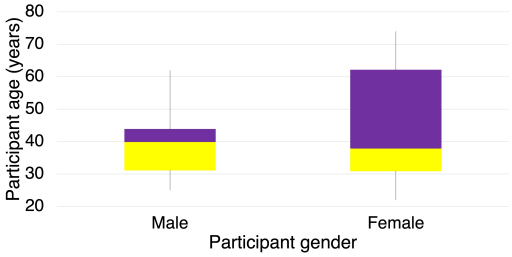

Figure 4.10: Box plot of participant age distribution by gender

rules in Figure 4.11. 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

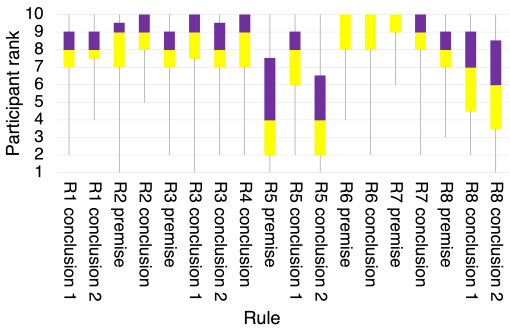

Figure 4.11: Box plot of participant rule rank

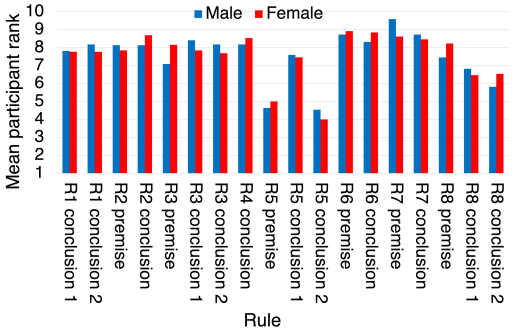

Figure 4.12: Bar plot of average rule rank by gender

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 R3 premise rule, the R3 conclusion 2 rule, the R5 conclusion 1 rule and the R8 premise rule, share a median of 8, with 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 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 4.12. For each rule, the average rank was computed for male and female participants respectively. According to this separation, our presentation of results in Figure 4.12 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. We show the contingency tables with the spread of the endorsement of the premises and conclusion for each AGM postulate in Tables 4.10–4.18. Each contingency

table consists of 2 rows and 2 columns. 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

Table 4.10: R1 (i)

31	4
0	0

Table 4.11: R1 (ii)

34	1
0	0

Table 4.12: R2

31	1
2	1

Table 4.13: R3

26	2
7	0

Table 4.14: R4 (i)

30	5
0	0

Table 4.15: R5

4	19
5	17

Table 4.16: R6

32	1
0	2

Table 4.17: R7

31	4
0	0

Table 4.18: R8

28	1
4	2

has been decomposed into two smaller postulates. Consequently, each part has its own premises 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. Considering all postulates, strict violations defined as endorsements of the premises but not the conclusion (B cells) were only 8,89%.

$$\text{cell B endorsements} = \frac{\text{no. of B responses across all postulates}}{315} * 100 \quad (4.3)$$

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 the premises and conclusion (A cells) were 78,41%. Non-endorsements of both the premises and conclusion were 6,98%. Non-endorsements of the premises and endorsements of the conclusion (C cells) were 5,71%. As a whole, endorsement rates of the premises were as expected.

Looking at the AGM postulates, the ϕ degrees of association reported in Table 4.19 show a highly significant association between the premises and

Table 4.19: 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 5% level. "$*$" refers to the vanilla two-tailed chi-squared test."$" refers to the two-tailed chi-squared test with Yates' correction. "†" refers to 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	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

conclusion postulate R6, using the two-tailed chi-squared test, chi-squared test with Yates' correction and Fisher's Exact Test. The ϕ degrees of association for postulates R2 and R8 are only statistically significant using the two-tailed unmodified chi-squared 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 4.19 also show a non-significant association between the premises and conclusion for postulates R3 and R5. The significance between the premises and conclusion for postulates R1 (i), R1 (ii), R4 (i) and R7 could not be determined because the conditions of the chi-squared 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 the conclusion of R2, R6 and R8. Considering all postulates, we tallied how many logical violations (out of 9 possible) each participant made. Results appear in Table 4.20. 45,71% made no violation at all, and 82,86% made one violation or none. These percentages are significantly high if we consider the unavoidable imperfection of our experimental apparatus, material and participant sampling. These results suggest that most participants tend to draw inferences from their knowledge that is consistent with the AGM postulates. In summary, we have shown 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 4.20: 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

Finally, a proportion of 45,71% of participants made no violation at all, and 82,86% made one violation or none, as shown in Table 4.20. These percentages suggest that most participants draw inferences from their knowledge in a manner that is consistent with the AGM belief revision postulates.

4.4 Experiment 4

In this experiment, we used the same material from the belief revision experiment to instantiate the KM belief update postulates. The experimental setup follows a similar approach to the belief revision experiment.

4.4.1 Participants

50 English-speaking participants were recruited on MTurk for this experiment.

4.4.2 Material

The postulates of belief update tested in this experiment are the 9 KM belief update postulates: *U1*, *U2*, *U3*,*U4*, *U5*, *U6*, *U7*, *U8* and *U9*. Each postulate was decomposed into its premises and conclusion. We refer to these components as rules. We show the rules for each of each postulate in Tables 4.21–4.29. Altogether, there are 22 rules consisting of either premise rules or conclusion rules. As before, we excluded rules that are tautologies from our survey and assumed that our participants would find tautologies plausible. The next step was to translate each rule into English. The same statements

selected for Experiment 3 were used in the translation. This was done to have a comparable knowledge base over the two belief change experiments. Once instantiated, the rules were included in the survey and the task was for each participant to rank the plausibility of the concrete rule on a scale from 1 (implausible) to 10 (extremely plausible). We divided the 22 rules between two surveys with an equal but smaller number of rules. We show the survey material for this experiment in Appendix B in Figures B.11–B.15. The survey URLs are provided in Appendix A.

Table 4.21: Decomposition of Postulate U1

Postulate	$K \diamond \mu \models \mu$
Material implication rule	If $\top$ then $K \diamond \mu \models \mu$
Premise	$\top$
Conclusion	$K \diamond \mu \models \mu$

Table 4.22: Decomposition of Postulate U2

Postulate	If $K \models \mu$ then $K \diamond \mu$ iff K
Material implication rule	If $K \models \mu$ then $K \diamond \mu$ iff K
Premise rule	$K \models \mu$
Conclusion rule 1	If $K \diamond \mu$ then K
Conclusion rule 2	If K then $K \diamond \mu$

Table 4.23: Decomposition of Postulate U3

Postulate	If both K and μ is satisfiable then $K \diamond \mu$ is satisfiable
Material implication rule	If both K and μ is satisfiable then $K \diamond \mu$ is satisfiable
Premise rule	both K and μ is satisfiable
Conclusion rule	$K \diamond \mu$ is satisfiable

4.4.3 Design and Procedure

Participants were recruited on MTurk and we established an internal set of criteria that they needed to satisfy. The criteria were identical to the belief revision experiment. The criteria set on MTurk for the first survey, testing

Table 4.24: Decomposition of Postulate U4

Postulate	If K_1 iff K_2 and μ_1 iff μ_2 then $K_1 \diamond \mu_1$ iff $K_2 \diamond \mu_2$
Material implication rule	If K_1 iff K_2 and μ_1 iff μ_2 then $K_1 \diamond \mu_1$ iff $K_2 \diamond \mu_2$
Premise rule 1	If K_1 then K_2
Premise rule 2	If K_2 then K_1
Premise rule 3	If μ_1 then μ_2
Premise rule 4	If μ_2 then μ_1
Conclusion rule 1	If $K_1 \diamond \mu_1$ then $K_2 \diamond \mu_2$
Conclusion rule 2	If $K_2 \diamond \mu_2$ then $K_1 \diamond \mu_1$

Table 4.25: Decomposition of Postulate U5

Postulate	$(\psi \diamond \mu) \wedge \phi \models \psi \diamond (\mu \wedge \phi)$
Material implication rule	If $\top$ then $(K \diamond \mu) \wedge \phi \models K \diamond (\mu \wedge \phi)$
Premise rule	$\top$
Conclusion rule	$(K \diamond \mu) \wedge \phi \models K \diamond (\mu \wedge \phi)$

the concrete rules for KM belief update postulates U1 to U5, are shown in Figure 4.13. Participants were not required to be Master workers and the *Task Visibility* requirement was set to *Hidden*. The criteria set on MTurk for the second survey, testing the concrete rules for KM belief update postulates U6 to U9, are shown in Figure 4.14. These criteria were identical to the first part. However, as with the previous experiment, we needed a manner in which to grant access only to those participants who had completed the first survey. To do this, we created two new qualification types on MTurk: *Completed_Part_1* and *Completed_Part_2*. Each qualification type also has an associated numeric score, between 0 and 100, that could be assigned to Workers. We assigned the qualification types *Completed_Part_1* with score 100 and *Completed_Part_2* with score 0 to indicate that a specific Worker has successfully completed the first survey, but has not completed the second survey. Upon completing the second survey, we updated a participant's *Completed_Part_2* score to 100. In addition to the Worker requirements, our survey assignment on MTurk was published after following the same project creation steps as the belief revision experiment. For the survey settings part, we specified the fields *Reward per response* as \$3,75, the *Number of respondents* as 50, *Time allotted per Worker* as 1 hour, *Survey expires in* as 7 days and *Auto-approve and pay Workers in* as 3 days. We show screenshots of

Table 4.26: Decomposition of Postulate U6

Postulate	If $K \diamond \mu_1 \models \mu_2$ and $K \diamond \mu_2 \models \mu_1$ then $K \diamond \mu_1$ iff $K \diamond \mu_2$
Material implication rule	If $K \diamond \mu_1 \models \mu_2$ and $K \diamond \mu_2 \models \mu_1$ then $K \diamond \mu_1$ iff $K \diamond \mu_2$
Premise rule	$K \diamond \mu_1 \models \mu_2$ and $K \diamond \mu_2 \models \mu_1$
Conclusion rule 1	If $K \diamond \mu_1$ then $K \diamond \mu_2$
Conclusion rule 2	If $K \diamond \mu_2$ then $K \diamond \mu_1$

Table 4.27: Decomposition of Postulate U7

Postulate	If K is complete then $(K \diamond \mu_1) \wedge (K \diamond \mu_2) \models K \diamond (\mu_1 \vee \mu_2)$
Material implication rule	If K is complete then $(K \diamond \mu_1) \wedge (K \diamond \mu_2) \models K \diamond (\mu_1 \vee \mu_2)$
Premise rule	K is complete
Conclusion rule	$(K \diamond \mu_1) \wedge (K \diamond \mu_2) \models K \diamond (\mu_1 \vee \mu_2)$

these settings from MTurk in Figures 4.13–4.14. The process of creating a new project on MTurk was applied twice, one for the survey testing KM postulates U1 to U5 and one for the survey testing KM postulates U6 to U9. To participate in the second part of the survey testing belief update, our participants needed to remember the embedded access code from the first survey, in addition to meeting the worker requirements on MTurk. During the data collection period, several eligible participants have misplaced or forgotten the access code for the second survey. While they were allowed to open the HIT by virtue of meeting the criteria on MTurk, it was found that the HIT had expired by the time they had contacted us with the issue. This led us to create a new qualification type on MTurk: *Participation_Round_2*. We rejected the incomplete, expired HITs and reissued HITs for participants who had difficulty with the access code. We assigned the qualification type *Participation_Round_2* with a score of 0 to Workers who were eligible but had not completed the second survey due to difficulty with the access code. Once these Workers had completed the second survey, we updated their *Participation_Round_2* score to 0. Participants were allowed to participate in the second survey as soon as they have completed the first. The participant procedures for the belief update surveys were similar to the belief revision surveys. We describe it in the following.

Upon accepting the HIT and gaining access to our first survey, participants were asked to authenticate themselves. This was done through an image-based trick question. A screenshot of this question is included in

Table 4.28: Decomposition of Postulate U8

Postulate	$(K_1 \vee K_2) \diamond \mu$ iff $(K_1 \diamond \mu) \vee (K_2 \diamond \mu)$
Material implication rule	If $\top$ then $(K_1 \vee K_2) \diamond \mu$ iff $(K_1 \diamond \mu) \vee (K_2 \diamond \mu)$
Premise rule	$\top$
Conclusion rule 1	If $(K_1 \vee K_2) \diamond \mu$ then $(K_1 \diamond \mu) \vee (K_2 \diamond \mu)$
Conclusion rule 2	If $(K_1 \diamond \mu) \vee (K_2 \diamond \mu)$ then $(K_1 \vee K_2) \diamond \mu$

Table 4.29: Decomposition of Postulate U9

Postulate	If K is complete and $(K \diamond \mu) \wedge \phi$ is satisfiable then $K \diamond (\mu \wedge \phi) \models (K \diamond \mu) \wedge \phi$
Material implication rule	If K is complete and $(K \diamond \mu) \wedge \phi$ is satisfiable then $K \diamond (\mu \wedge \phi) \models (K \diamond \mu) \wedge \phi$
Premise rule	K is complete and $(K \diamond \mu) \wedge \phi$ is satisfiable
Conclusion rule	$K \diamond (\mu \wedge \phi) \models (K \diamond \mu) \wedge \phi$

Figure 4.15. This question was introduced to prove that participants were human. Upon successfully identifying the purple balloons among a set of images including red, pink and blue balloons, participants were allowed to continue with the survey. Upon selecting the incorrect message, a feedback window was displayed and participants were allowed to redo the authentication question. By managing the authentication this way, we encouraged our participants to be alert and pay attention to their responses. The next part of the survey was the consent form which included an invitation to participate, information about procedures, recording, risks and feedback, a disclaimer, a confidentiality agreement and information about what signing the consent form means. The next part of the survey asked participants to confirm that they satisfy the survey inclusion criteria, that they 1) are 18 years old or above, 2) reside in the United States of America and 3) are fluent in English. The next part of the survey asked participants to provide their age and gender. The next part of the survey listed a set of key words and their explanation, introducing the topic matter of the survey. Participants were advised to read the key words carefully before proceeding with the rest of the survey. The key words are included in the project repository, linked in Appendix C. As before, at any point in the survey, participants were allowed to navigate back to the key words during the survey. The key words were also placed in a Google Doc and linked as a URL in a text area above each

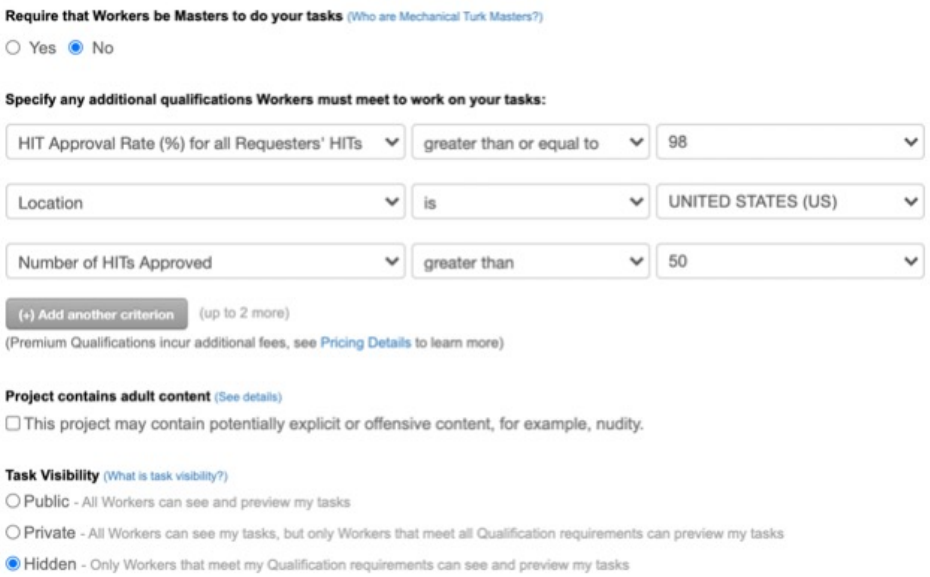

Figure 4.13: Worker requirements on MTurk for the KM belief update survey testing postulates U1 to U5

rule, for ease of access to the key words. Next, participants were provided with a practice question. Passing the practice question required participants to select a value on a linear 10-point scale, from 1 = extremely implausible to 10 = plausible, corresponding to how plausible they find the given rule. For the practice question, participants were provided with a feedback prompt indicating success if they had selected a value on the scale from 1 to 10 corresponding to how plausible they found the rule in the example. Each rule in the survey was marked as *required (*)*. In the main part of the survey, participants were provided with the concrete rules for KM belief update postulates U1 to U5 and were asked to rate its plausibility, as in the practice question. The main part of the survey was estimated to take 10 to 20 minutes at a leisurely pace. At the end of the survey, participants were provided with a unique survey code that they needed to enter on MTurk to receive payment. Participants were informed that they needed to provide their MTurk WorkerID before submitting the survey. There was no bonus reward for this survey. At this point in the survey, participants were also given a survey access code to the second survey. Although participants were given 60 minutes to complete the first survey, it was expected that they might com-

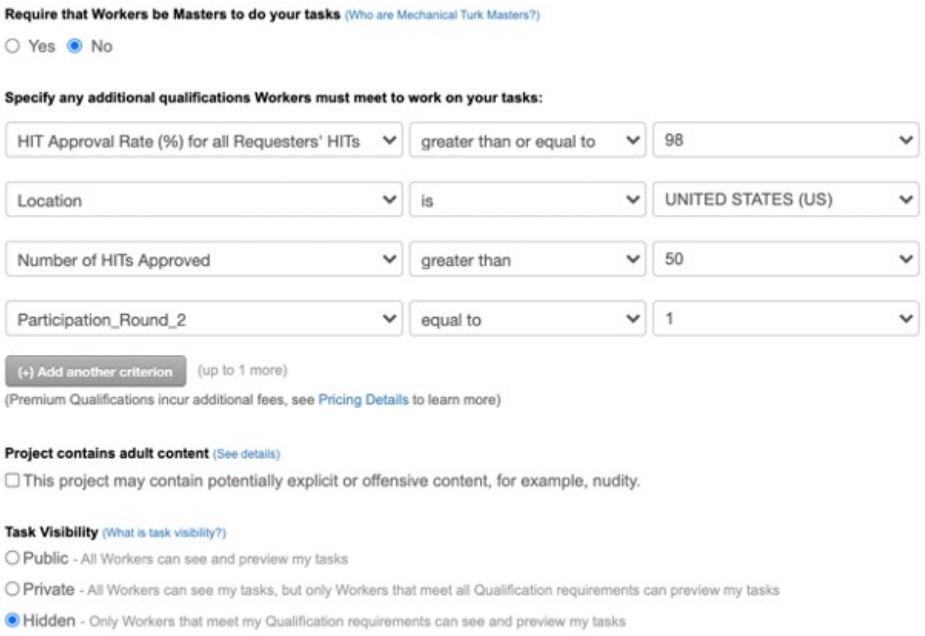

Figure 4.14: Worker requirements on MTurk for the KM belief update survey testing postulates U6 to U9

plete it within 20 to 40 minutes. At the end of the survey, participants were thanked for their time and allowed to provide feedback about their experience. The second part of the survey followed the same design and procedure as the first part, except for a few differences. The authentication question asked participants to identify a smartphone in a set of images also containing a wired telephone, a gramophone and a radio. A screenshot of this question is included in Figure 4.16. No practice question was included in the second survey. The main part of the second survey contained the concrete rules for KM belief update postulates U6 to U9.

4.4.4 Predictions

We aimed to determine whether the endorsement of the premises was preferentially associated with the endorsement of the conclusion of each postulate. We computed a contingency table that corresponds to the combinations of endorsement of the premises and conclusion in classical logic. To identify in which cell of the contingency table a participant's response matched, we

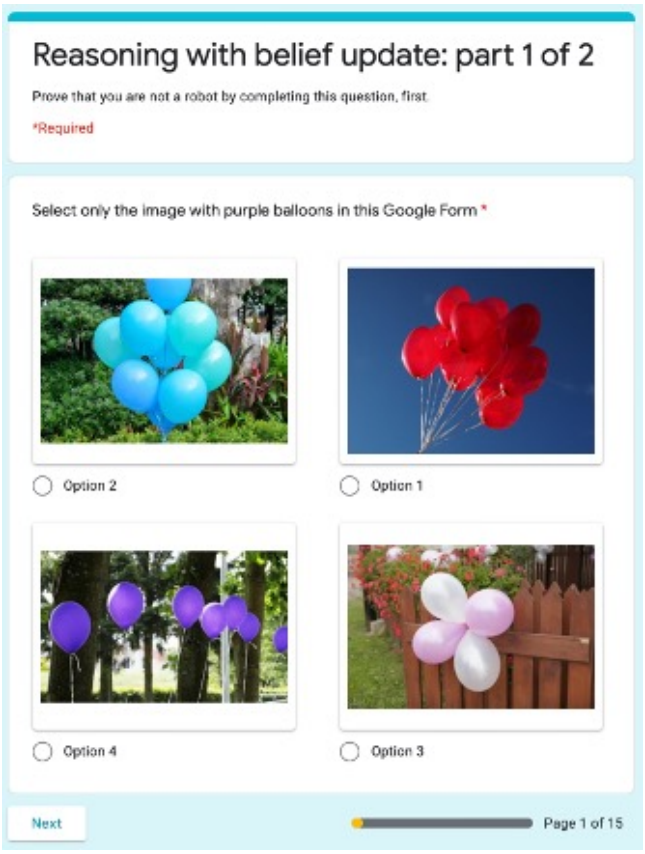

Figure 4.15: Participant authentication question for KM belief update survey testing postulates U1 to U5

applied the logical evaluation of the implication, premises $\longrightarrow$ conclusion. The degree of association between the endorsement of the premises and the endorsement of the conclusion was computed using the phi-coefficient. For each postulate, the statistical hypothesis H_0 (null hypothesis) was "there is no significant association between the endorsement of the premises and the endorsement of the conclusion". The alternate hypothesis, H_1, was "there is a significant positive association between the endorsement of the premises and the endorsement of the conclusion". 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. Under the general hypothesis that the participants' inference tends to corroborate the 9 KM postulates, we predicted that H_0 would be rejected for each postulate. Under the hypothesis that participant judgements are consistent with

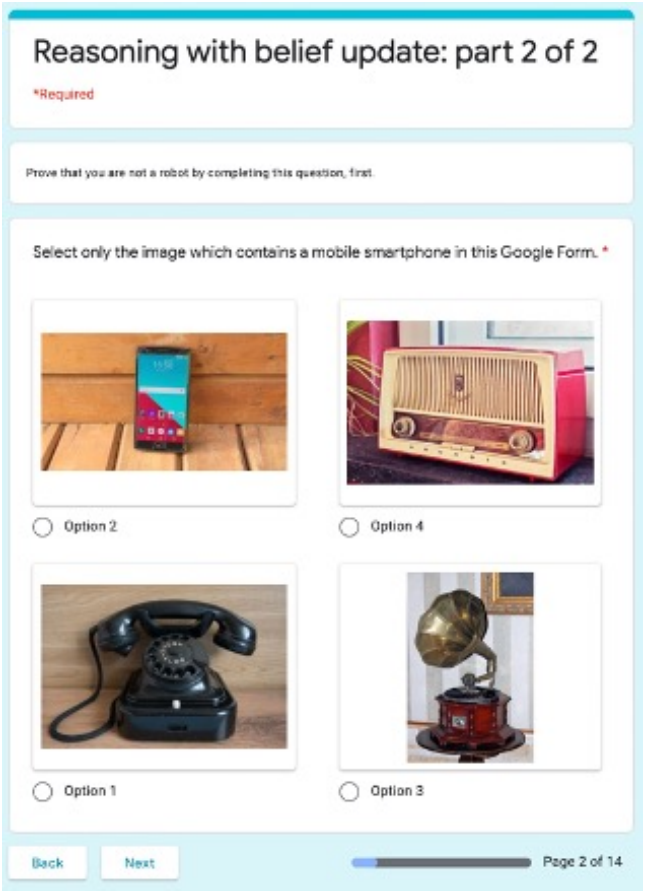

Figure 4.16: Participant authentication question for KM belief update survey testing postulates U6 to U9

the KM postulates, we predicted a high proportion of participants would not commit any logical violation.

4.4.5 Experimental Results

We downloaded the survey responses from Google Forms and used Microsoft Excel for coding. The rank given by each participant was coded as either True or False. The code True was assigned to statements with a plausibility rating of > 5 and ≤ 10. The code False was assigned to statements with a plausibility rating of ≥ 1 and < 6. Where premise or conclusion rules contain material implication or material equivalence statements, for example,

postulates U2, U4, U6 and U8, these rules were decomposed further. The decomposition process was applied until the premise rule and conclusion rule for each postulate involved only a single material implication statement. Each rule obtained by this decomposition process was tested in the survey. In our coding process, rules were grouped according to the part of the postulate that it represented, the premises or the conclusion. The overall value for the premises of a postulate was obtained by the logical valuation of the conjunction of all the codes used in the rules for the premises. The same process was used to obtain the overall value for the conclusion. In the case of the premise 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 the conclusion. During the data collection process, 50 participants completed the first belief revision survey. Although each participant expressed interest in taking the second belief revision survey, only 37 participants responded to the second survey. For a comparable dataset to the belief revision experiment, we further removed two random complete responses. As a result, our adjusted sample size for each of the surveys is 35 participants. Once the data was collected, a data cleaning process was performed. This step was programmed in R and the script is available in our project repository, linked in Appendix C. An inspection of the data from MTurk revealed that the *Lifetime Approval Rate for Requester's 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 5. This indicates how familiar Workers were 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 revision, with 8,11% (3 Workers) having taken 1 of our tasks previously, an additional 8,11% having taken 2 of our tasks previously and a further 2,7% (1 Worker) 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 4.17. Participants ages ranged from 24 to 61 years old. In Figure 4.18, 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 the participant rank for

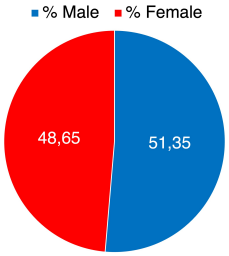

Figure 4.17: Pie chart of participant gender distribution by percentage

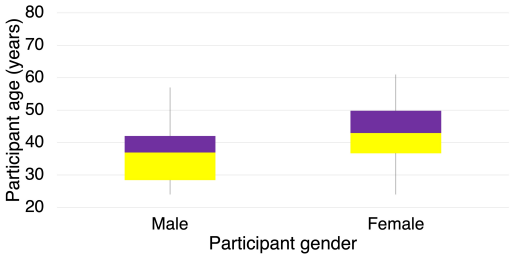

Figure 4.18: Box plot of participant age distribution by gender

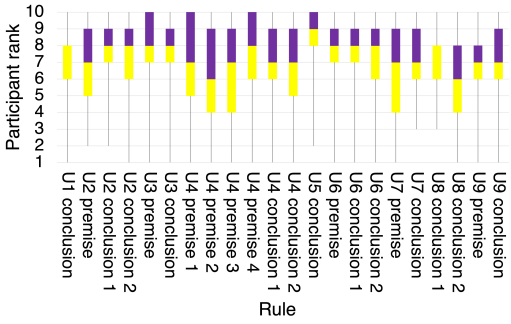

Figure 4.19: Box plot of participant rule rank

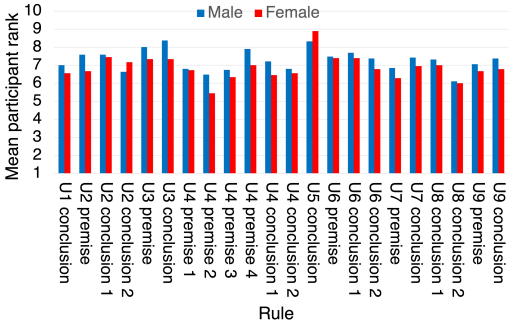

Figure 4.20: Bar plot of average rule rank by gender

our 22 concrete rules in Figure 4.19. 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 rank of the premises of postulates U2, U4 and U7, suggest that those postulates have contentious premises with ranks also ranging from moderately low 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. The box plots for the U2 conclusion 1 rule, U4 premise 1 rule, U6 premise rule, U6 conclusion 1 rule and U9 premise rule are short and even in size in the second and third quartiles, suggesting responses are evenly distributed around their median rank. The box plots for the U2 premise rule, the U4 premise 4 rule and the U8 conclusion 2 rule, are even in size in the second and third quartiles, but are comparatively taller than the other box plots with even distributions in the second and third quartiles. This suggests that even though responses are evenly distributed around their median rank, there is a greater range of responses for these rules. The bar plots for the U2 conclusion 2 rule, U4 premise 3 rule, U6 conclusion 2 rule and U7 premise rules are skewed to the left of the median. The bar plots for the U3 premise rule, U4 premise 1 rule, U4 premise 2 rule, U7 conclusion rule and U9 conclusion rule are skewed to the right of the median. 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 4.20. For each rule, the average rank was computed for male and female participants respectively. According to this separation, our presentation of results in Figure 4.20 should not be biased because the number of male and female participants only differ by 1 member. Additionally, we investigated 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 differs for male and female participants by only one rank. 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.

To determine the effect of the endorsement of the premises and the en-

dorsement of the conclusion of each postulate. We did this through contingency tables, as with the belief revision experiment. We show the contingency tables with the spread of the endorsement of the premises and the conclusion of each KM postulates in Tables 4.30–4.38. 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

Table 4.30: U1

29	8
0	0

Table 4.31: U2

25	1
4	7

Table 4.32: U3

31	0
4	2

Table 4.33: U4

13	0
7	17

Table 4.34: U5

35	2
0	0

Table 4.35: U6

24	5
0	8

Table 4.36: U7

21	4
7	5

Table 4.37: U8

20	17
0	0

Table 4.38: U9

25	5
4	4

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 were only 12,61%.

$$\text{cell B endorsements} = \frac{\text{no. of B responses across all postulates}}{333} * 100 \quad (4.4)$$

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 the premises and conclusion (A cells) were 66,67%. Non-endorsements of both the premises and conclusion (D cells) were 12,91%. Non-endorsements of the premises and endorsements of the conclusion (C cells) were 7,81%. As a whole, endorsement rates of the premises were as expected.
Looking at the KM belief update postulates, the ϕ degrees of association

Table 4.39: 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-squared test. "$" refers to the two-tailed chi-squared 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

reported in Table 4.39 show a highly significant association between the premises and the conclusion for postulates U2, U4 and U6, and a weakly significant association between the premises and conclusion for postulate U3, using the two-tailed chi-squared test, chi-squared 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-squared 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 4.39 also show a non-significant association between the premises and the conclusion for postulates U1, U5 and U8, using Fisher's Exact Test. Postulate U7 has a non-significant association between the premises and conclusion in the cases of the chi-squared test with Yates' correction and Fisher's Exact Test, but not the vanilla chi-squared test. U9 has a non-significant association between the premises and conclusion only in the case of the chi-squared test with Yates' correction. The significance between the premises and conclusion for postulates U1, U5 and U8, using the vanilla chi-squared test, could not be determined because the conditions of the chi-squared 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 the 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 4.40. 18,92% of participants made no violation at all, and 70,27% made one violation or none. These percentages are significantly high if we consider the unavoidable imperfection of our experimental apparatus, material, and participant sampling. These results suggest that most participants tend to draw inferences from their knowledge that is consistent with the KM belief update postulates. In summary, we have shown that postulates U2, U3, R4, U6, U7 and U9 were significantly endorsed by our participants. The results for Postulates U1, U5 and U8 are inconclusive, according to our criteria. Finally, a proportion

Table 4.40: 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

of 18,92% of participants made no violation at all, and 70,27% made one violation or none, as shown in Table 4.40. These percentages suggest that most participants draw inferences from their knowledge in a manner that is consistent with the KM belief update postulates.

Chapter 5

Results and Discussion

In this chapter, we focus on the results from the belief revision belief update experiments. We discuss how these results relate to our research hypothesis and contribute to the literature on belief change in the cognitive science, psychology and AI communities.

5.1 Human Reasoning and the AGM Postulates

In our survey testing the AGM postulates, we have introduced new terminology for reasoning with belief revision. We defined the terms "consequences", "contained in your beliefs", "equivalent", "expansion", "includes", "interpretation", "K (knowledge base)", "new information", "revision" and "satisfiable". This was done to ensure that all our participants had a common understanding of these terms and their usage. In the following, we discuss the belief revision results at a postulate level. R1 (i) (*Closure*) has a non-significant association between its premises and conclusion, according to Fisher's exact test. 88,57% (31 participants) endorsed both the premise rule and the conclusion rule. 11,43% (4 participants) committed a violation by endorsing the premises, but not the conclusion. No participants endorsed the conclusion while not endorsing the premises. No participants endorsed either the premises or the conclusion. In our decomposition of postulate R1 (i), we obtained one premises rule and one conclusion rule. The premises rule for R1 (i) is a tautology, $\top$. The conclusion rule for postulate R1 (i) states that K includes not only the beliefs contained in it, but also the consequences which

follow from K. According to the representation of default rules in possibility theory [10], a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R1 (i), we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R1 (i). It follows that $\Pi(\alpha \wedge \beta) = 31$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for R1 (i). It follows that $\Pi(\alpha \wedge \neg\beta) = 4$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\top \rightarrow \beta$ is found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human reasoning is consistent with AGM postulate R1 (i) is proven true for our participants, but not all human reasoners.

R1 (ii) (*Closure*) has a non-significant association between its premises and conclusion, according to Fisher's exact test. 97,14% (34 participants) endorsed both the premises and the conclusion. 2,86% (1 participant) committed a violation by endorsing the premise, but not the conclusion. No participants did not endorse the premises while endorsing the conclusion. No participants endorsed either the premises or the conclusion. In our decomposition of postulate R1 (ii), we obtained one premises rule and one conclusion rule. The premises rule for R1 (ii) is a tautology, $\top$. The conclusion rule for R1 (ii) states that the result of revising K with the new information that Zeeta M is a classical pianist also contains the information which follows from the result of revising K with the new information that Zeeta M is a classical pianist. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R1 (ii), we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R1(ii). It follows that $\Pi(\alpha \wedge \beta) = 34$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for R1 (ii). It follows that $\Pi(\alpha \wedge \neg\beta) = 1$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\top \rightarrow \beta$ is found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that

this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human reasoning is consistent with AGM postulate R1 (ii) is proven true for our participants, but not all human reasoners.

R2 (*Inclusion*) has a significant association between its premises and conclusion, according to the chi-squared test. 88,57% (31 participants) endorsed both the premises and the conclusion. 2,86% (1 participant) committed a violation by endorsing the premises, but not the conclusion. 5,71% (2 participants) did not endorse the premises, but did endorse the conclusion. 2,86% (1 participant) endorsed neither the premises nor the conclusion. In our decomposition of postulate R2, we obtained one premises rule and one conclusion rule. The premise rule for R2 states that Jacob B does drive at night follows from revising K with the new information that Jacob B is a truck driver. The conclusion rule for R2 states that Jacob B does drive at night follows from the expanding K with the new information that Jacob B is a truck driver. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R2, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R2. It follows that $\Pi(\alpha \wedge \beta) = 31$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for R2. It follows that $\Pi(\alpha \wedge \neg\beta) = 1$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\alpha \rightarrow \beta$ is found plausible by our participants. Since the association between the premises and conclusion is statistically significant, this result generalises to all English-speaking human reasoners. Our hypothesis that human reasoning is consistent with AGM postulate R2 is thus proven true for our participants and it generalises to all English-speaking human reasoners.

R3 (*Vacuity*) has a non-significant association between its premises and conclusion, according to the chi-squared test, chi-squared test with Yates' correction and Fisher's exact test. 74,29% (26 participants) endorsed both the premises and the conclusion. 5,71% (2 participants) committed a violation by endorsing the premises, but not the conclusion. 20% (7 participants) did not endorse the premises, but did endorse the conclusion. No participants endorsed either the premises or the conclusion. In our decomposition postulate R3, we obtained one premises rule and two conclusion rules. The

premises rule for R3 states that K is satisfiable with respect to the new information that Jessica B is a yoga instructor. The first conclusion rule for R3 states that Jessica B does teach breathing exercises follows from the expansion of K with the new information that Jessica B is a yoga instructor. The second conclusion rule for R3 states that Jessica B does teach breathing exercises follows from the revision of K with the new information that Jessica B is a yoga instructor. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R3, we let α represent our premises rule and we let $\beta \wedge \gamma$ represent the conjunction of our conclusion rules, β and γ. $\Pi(\alpha \wedge (\beta \wedge \gamma))$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R2. It follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) = 26$. $\Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for R2. It follows that $\Pi(\alpha \wedge \neg(\beta \wedge \gamma)) = 2$. Further, it follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) > \Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ and that a rule of the form $\alpha \rightarrow (\beta \wedge \gamma)$ is found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human is consistent with AGM postulate R3 is proven true for our participants, but not all human reasoners.

R4 (*Success*) has a non-significant association between its premises and conclusion, according to Fisher's exact test. 85,71% (30 participants) endorsed both the premises and the conclusion. 14,29% (5 participants) committed a violation by endorsing the premises, but not the conclusion rule. No participants did not endorse the premises while endorsing the conclusion. No participants endorsed either the premises or the conclusion. In our decomposition of postulate R4, we obtained one premises rule and one conclusion rule. The premises rule for R4 is a tautology, $\top$. The conclusion rule for R4 states that the new information that Chris P is a waiter is contained in the revision of K with the new information that Chris P is a waiter. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R4, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R4. It follows that $\Pi(\alpha \wedge \beta) = 30$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did

not endorse the conclusion, seen in Cell B of the contingency table for R4. It follows that $\Pi(\alpha \wedge \neg\beta) = 4$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\top \rightarrow \beta$ is found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human reasoning is consistent with AGM postulate R4 is proven true for our participants, but not all human reasoners.

R5 (*Extensionality*) has a non-significant association between its premises and conclusion, according to the chi-squared test, chi-squared test with Yates' correction and Fisher's exact test. 11,43% (4 participants) endorsed both the premises and the conclusion. 54,29% (19 participants) committed a violation by endorsing the premises, but not the conclusion. 14,29% (5 participants) did not endorse the premises, but did endorse the conclusion. 48,57% (17 participants endorsed neither the premises nor the conclusion. In our decomposition of postulate R5, we obtained one premises rule and two conclusion rules. The premise rule for R5 states that the new information that if Noel W is a firefighter then Noel W is strong is equivalent to the new information that either Noel W is not a firefighter or Noel W is strong. The first conclusion rule for R5 states that Noel W does save lives follows from the revision of K with the new information that Noel W is a strong firefighter. The second conclusion rule for R5 states that Noel W does save lives follows from the revision of K with the new information that either Noel W is not a firefighter or Noel W is strong. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R5, we let α represent our premises rule and we let $\beta \wedge \gamma$ represent the conjunction of our conclusion rules, β and γ. $\Pi(\alpha \wedge (\beta \wedge \gamma))$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R5. It follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) = 4$. $\Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for R5. It follows that $\Pi(\alpha \wedge \neg(\beta \wedge \gamma)) = 19$. Further, it follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) \leq \Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ and that a rule of the form $\alpha \rightarrow (\beta \wedge \gamma)$ is not found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking

human reasoners. Thus, our hypothesis that human reasoning is consistent with the AGM postulates R5 is proven false for our participants, but not all human reasoners.

R6 (*Consistency*) has a highly significant association between its premises and conclusion, according to all three statistical tests: chi-squared, chi-squared with Yates' correction and Fisher's exact test. 91,43% (32 participants) endorsed both the premises and the conclusion. 2,86% (1 participant) committed a violation by endorsing the premises, but not the conclusion. No participants did not endorse the premises while endorsing the conclusion. 5,71% (2 participants) endorsed neither the premises nor the conclusion. In our decomposition of postulate R6, we obtained one premises rule and one conclusion rule. The premises rule for R6 states that the new information that Wilma D is a car owner is consistent. The conclusion rule states that the revision of K with the new information that Wilma D is a car owner is consistent. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R6, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R6. It follows that $\Pi(\alpha \wedge \beta) = 32$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for R6. It follows that $\Pi(\alpha \wedge \neg\beta) = 1$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\alpha \rightarrow \beta$ is found plausible by our participants. Since the association between the premises and conclusion is highly statistically significant, this result generalises to all English-speaking human reasoners. Our hypothesis that human reasoning is consistent with AGM postulate R6 is thus proven true for our participants and it generalises to all English-speaking human reasoners.

R7 (*Super-expansion*) has a non-significant association between its premises and conclusion, according to Fisher's exact test. 88,57% (31 participants) endorsed both the premises and the conclusion. 11,43% (4 participants) committed a violation by endorsing the premises, but not the conclusion. No participants did not endorse the premises while endorsing the conclusion. No participants endorsed either the premises or the conclusion. In our decomposition of postulate R7, we obtained one premises rule and one conclusion rule. The premises rule for R7 states that Phillip P does carry a gun follows from the revision of K with the new information that Phillip P is a police officer and Phillip P can arrest a criminal. The conclusion rule for R7 states

that Phillip P does carry a gun follows from the result of first revising K with the new information that Phillip P is a police officer and then expanding with the new information that Phillip P can arrest a criminal. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R7, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R7. It follows that $\Pi(\alpha \wedge \beta) = 31$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for R7. It follows that $\Pi(\alpha \wedge \neg\beta) = 4$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\alpha \rightarrow \beta$ is found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human reasoning is consistent with AGM postulate R7 is proven true for our participants, but not all human reasoners.

R8 (*Super-expansion*) has a significant association between its premises and conclusion, according to the chi-squared test. 80% (28 participants) endorsed both the premises and the conclusion. 2,86% (1 participant) committed a violation by endorsing the premises, but not the conclusion. 11,43% (4 participants) did not endorse the premises, but did endorse the conclusion. 5,71% (2 participants) endorsed neither the premises nor the conclusion. In our decomposition of postulate R8, we obtained one premises rule and two conclusion rules. The premise rule for R8 states that the revision of K with the new information that Mark M is a science professor is satisfiable with respect to the new information that Mark M does enjoy solving problems. The first conclusion rule for R8 states that Mark M is a good teacher follows from the result of revising K with the new information that Mark M is a science professor and then expanding with the new information that Mark M does enjoy solving problems. The second conclusion rule for R8 states that Mark M is a good teacher follows from the result of revising K with the new information that Mark M is a science professor and Mark M does enjoy solving problems. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate R5, we let α represent our premises rule and we let $\beta \wedge \gamma$ represent the conjunction of our conclusion rules, β and γ. $\Pi(\alpha \wedge (\beta \wedge \gamma))$ is the number

of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for R8. It follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) = 28$. $\Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for R8. It follows that $\Pi(\alpha \wedge \neg(\beta \wedge \gamma)) = 1$. Further, it follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) > \Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ and that a rule of the form $\alpha \rightarrow (\beta \wedge \gamma)$ is found plausible by our participants. Since the association between the premises and conclusion is statistically significant, this result generalises to all English-speaking human reasoners. Our hypothesis that human reasoning is consistent with AGM postulate R8 is thus proven true for our participants and it generalises to all English-speaking human reasoners.

5.2 Human Reasoning and the KM Postulates

In our survey testing the KM postulates, we have introduced new terminology for reasoning with belief update. We defined the terms "Ki (belief set)", "$K_1 \vee K_2$ (disjunction of two belief bases)", "$K_1 \wedge K_2$ (conjunction of two belief sets)", "completeness", "consequences", "interpretation", "new information", "satisfiable" and "update". This was done to ensure that all our participants have a common understanding of these terms and their usage. In the following, we discuss the belief update results at a postulate level. U1 has a non-significant association between its premises and conclusion, according to Fisher's exact test. 78,38% (29 participants) endorsed both the premises and the conclusion. 21,62% (8 participants) committed a violation by endorsing the premises, but not the conclusion. No participants did not endorse the premises while endorsing the conclusion. No participants endorsed either the premises or the conclusion. In our decomposition of postulate U1, we have obtained one premises rule and one conclusion rule. The premises rule for U1 is a tautology, $\top$. The conclusion rule for U1 states that the new information that Jacob B is a truck driver follows from the result of updating K with the new information that Jacob B is a truck driver. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U1, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for U1. It follows that $\Pi(\alpha \wedge \beta) = 29$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did

not endorse the conclusion, seen in Cell B of the contingency table for U1. It follows that $\Pi(\alpha \wedge \neg\beta) = 8$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\top \rightarrow \beta$ is found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that this result is specific to our choice of material and there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human reasoning is consistent with KM belief update postulate U1 is proven true for our participants, but not for all human reasoners.

U2 has a highly significant association between its premises and conclusion, according to the chi-squared test, chi-squared test with Yates' correction and Fisher's exact test. 67,57% (25 participants) endorsed both the premises and the conclusion. 2,7% (1 participant) committed a violation by endorsing the premises, but not the conclusion. 10,81% (4 participants) did not endorse the premises while endorsing the conclusion. 18,92% (7 participants) endorsed neither the premises nor the conclusion. In our decomposition of postulate U2, we obtained one premises rule and two conclusion rules. The premise rule for U2 states that the new information that Noel W is a strong firefighter follows from K. The first conclusion rule states that K follows from the result of updating K with the new information that Noel W is a strong firefighter. The second conclusion rule for U2 states that the result of updating K with the new information that Noel W is a strong firefighter follows from K. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U2, we let α represent our premises rule and we let $\beta \wedge \gamma$ represent the conjunction of our conclusion rules, β and γ. $\Pi(\alpha \wedge (\beta \wedge \gamma))$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for U2. It follows that $\Pi(p \wedge (\beta \wedge \gamma)) = 25$. $\Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for U2. It follows that $\Pi(\alpha \wedge \neg(\beta \wedge \gamma)) = 1$. Further, it follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) > \Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ and that a rule of the form $\alpha \rightarrow (\beta \wedge \gamma)$ is found plausible by our participants. Since the association between the premises and conclusion is statistically significant, this result generalises to all English-speaking human reasoners. Our hypothesis that human reasoning is consistent with KM postulate U2 is thus proven true for our participants and it generalises to all English-speaking human reasoners.

U3 has a weakly significant association between its premises and conclu-

sion, according to the chi-squared test, chi-squared test with Yates' correction and Fisher's exact test. 83,78% (31 participants) endorsed both the premises and the conclusion. No participants committed a violation by endorsing the premises, but not the conclusion. 10,81% (4 participants) did not endorse the premises while endorsing the conclusion. 5,41% (2 participants) endorsed neither the premises nor the conclusion. In our decomposition of postulate U3, we obtained one premises rule and one conclusion rule. The premises rule for U3 states that both K and the new information that Eric V is a football player from Slovenia is satisfiable. The conclusion rule for U3 states that the result of updating K with the new information that Eric V is a football player from Slovenia is satisfiable. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U3, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for U3. It follows that $\Pi(\alpha \wedge \beta) = 31$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for U3. It follows that $\Pi(\alpha \wedge \neg\beta) = 0$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\alpha \rightarrow \beta$ is found plausible by our participants. Since the association between the premises and conclusion is statistically significant, albeit weakly, this result generalises to all English-speaking human reasoners. Our hypothesis that human reasoning is consistent with the KM postulate U3 is thus proven true for our participants and it generalises to all English-speaking human reasoners.

U4 has a highly significant association between its premises and conclusion, according to the chi-squared test, chi-squared test with Yates' correction and Fisher's exact test. 35,14% (13 participants) endorsed both the premises and the conclusion. No participants committed a violation by endorsing the premises, but not the conclusion. 18,92% (7 participants) did not endorse the premises while endorsing the conclusion. 45,95% (17 participants) endorsed neither the premises nor the conclusion. In our decomposition of postulate U4, we obtained four premises rules and two conclusion rules. The premises of postulate U4 is obtained by the conjunction of the individual premise rules. The conclusion is obtained by the conjunction of the individual conclusion rules. The results in our contingency table are for the overall premises and conclusion for postulate U4. The first premises rule for U4 states that K_2 follows from K_1. The second premises rule for U4 states that

K_1 follows from K_2. The third premises rule for U4 states that the new information that Zeeta M is a classical pianist follows from the new information that Zeeta M does play the piano. The fourth premises rule states that the new information that Zeeta M does play the piano follows from the new information that Zeeta M is a classical pianist. The first conclusion rule states that the result of updating K_2 with the new information that Zeeta M does play the piano follows from the result of updating K_1 with the new information that Zeeta M is a classical pianist. The second conclusion rule states that the result of updating K_1 with the new information that Zeeta M is a classical pianist follows from updating K_2 with the new information that Zeeta M does play the piano. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U2, we let $\alpha \wedge \epsilon \wedge \delta \wedge \rho$ represent the conjunction of our four premises rules and we let $\beta \wedge \gamma$ represent the conjunction of our conclusion rules. $\Pi((\alpha \wedge \epsilon \wedge \delta \wedge \rho) \wedge (\beta \wedge \gamma))$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for U4. It follows that $\Pi((\alpha \wedge \epsilon \wedge \delta \wedge \rho) \wedge (\beta \wedge \gamma)) = 13$. $\Pi((\alpha \wedge \epsilon \wedge \delta \wedge \rho) \wedge \neg(\beta \wedge \gamma))$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for U4. It follows that $\Pi((\alpha \wedge \epsilon \wedge \delta \wedge \rho) \wedge \neg(\beta \wedge \gamma)) = 0$. Further, it follows that $\Pi((\alpha \wedge \epsilon \wedge \delta \wedge \rho) \wedge (\beta \wedge \gamma)) > \Pi((\alpha \wedge \epsilon \wedge \delta \wedge \rho) \wedge \neg(\beta \wedge \gamma))$ and that a rule of the form $(\alpha \wedge \epsilon \wedge \delta \wedge \rho) \rightarrow (\beta \wedge \gamma)$ is found plausible by our participants. Since the association between the premises and conclusion is highly statistically significant, this result generalises to all English-speaking human reasoners. Our hypothesis human reasoning is consistent with the KM postulate U4 is proven true for our participants and generalises to all English-speaking human reasoners.

U5 has a non-significant association between its premises and conclusion according to Fisher's exact test. 94,59% (35 participants) endorsed both the premises and the conclusion. 5,41% (2 participants) committed a violation by endorsing the premises, but not the conclusion. No participants did not endorse the premises while endorsing the conclusion. No participants endorsed either the premises or the conclusion. In our decomposition of postulate U5, we obtained one premises rule and one conclusion rule. The premises rule for U5 is a tautology, $\top$. The conclusion rule for U5 states that the result of updating K_1 with the new information that Wilma D is a car owner $\wedge$ Wilma D does pay insurance follows from K_1. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible

iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U5, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for U5. It follows that $\Pi(\alpha \wedge \beta) = 35$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for U5. It follows that $\Pi(\alpha \wedge \neg\beta) = 2$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\top \rightarrow \beta$ is found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human reasoning is consistent with KM postulate U5 is proven true for our participants, but not for all human reasoners.

U6 has a highly significant association between its premises and conclusion, according to the chi-squared test, chi-squared test with Yates' correction and Fisher's exact test. 64,86% (24 participants) endorsed both the premises and the conclusion. 13,51% (5 participants) committed a violation by endorsing the premises, but not the conclusion. No participants did not endorse the premises while endorsing the conclusion. 21,62% (8 participants) endorsed neither the premises nor the conclusion. In our decomposition of postulate U6, we obtained one premises rule and two conclusion rules. The premises rule for U6 states that the new information that Chris P has profound knowledge of the menu follows from the result of updating K with the new information that Chris P is a waiter, and the new information that Chris P is a waiter follows from the result of updating K with the new information that Chris P has profound knowledge of the menu. The first conclusion rule for U6 states that the result of updating K with the new information that Chris P has profound knowledge of the menu follows from the result of updating K with the new information that Chris P is a waiter. The second conclusion rule for U6 states that the result of updating K with the new information that Chris P is a waiter follows from the result of updating K with the new information that Chris P has profound knowledge of the menu. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U2, we let α represent our premises rule and we let $\beta \wedge \gamma$ represent the conjunction of our conclusion rules, β and γ. $\Pi(\alpha \wedge (\beta \wedge \gamma))$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for

U6. It follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) = 24$. $\Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for U6. It follows that $\Pi(\alpha \wedge \neg(\beta \wedge \gamma)) = 5$. It is now clear that $\Pi(\alpha \wedge (\beta \wedge \gamma)) > \Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ and that a rule of the form $\alpha \rightarrow (\beta \wedge \gamma)$ is found plausible by our participants. Since the association between the premises and conclusion is highly statistically significant, this result generalises to all English-speaking human reasoners. Our hypothesis that human reasoning is consistent with KM postulates U6 is proven true for our participants and generalises to all English-speaking human reasoners.

U7 has a near significant association between its premises and conclusion, according to the chi-squared test, but a non-significant association according to the chi-squared test with Yates' correction and Fisher's exact test. 56,76% (21 participants) endorsed both the premises and the conclusion. 10,81% (4 participants) committed a violation by endorsing the premises, but not the conclusion. 18,92% (7 participants) did not endorse the premises while endorsing the conclusion. 13,51% (5 participants) endorsed neither the premises nor the conclusion. In our decomposition of postulate U7, we obtained one premises rule and one conclusion rule. The premises rule for U7 states that K is complete. The conclusion rule for U7 states that the result of updating K with the new information that Jessica B is a yoga instructor $\wedge$ Jessica B does teach breathing exercises follows from the result of updating K with the new information that Jessica B is a yoga instructor $\wedge$ the result of updating K with the new information that Jessica B does teach breathing exercises. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U5, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for U7. It follows that $\Pi(\alpha \wedge \beta) = 21$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for U7. It follows that $\Pi(\alpha \wedge \neg\beta) = 4$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\alpha \rightarrow \beta$ is found plausible by our participants. However, the association between the premises and conclusion is statistically non-significant to nearly significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human reasoning is consistent with KM postulate U7 is proven true for our participants, but not all human

reasoners.

U8 has a non-significant association between its premises and conclusion, according to Fisher's exact test. 54,05% (20 participants) endorsed both the premises and the conclusion. 45,95% (17 participants) committed a violation by endorsing the premises, but not the conclusion. No participants did not endorse the premises while endorsing the conclusion. No participants endorsed either the premises or the conclusion. In our decomposition of postulate U8, we obtained one premises rule and two conclusion rules. The premises rule for U8 is a tautology, $\top$. The first conclusion rule for U8 states that the result of updating K_1 with the new information that Phillip P is a police officer follows from the result of updating the new information $K_1 \vee K_2$ with the new information that Phillip P is a police officer. The second conclusion rule for U8 states that the result of updating the new information that $K_1 \vee K_2$ with the new information that Phillip p is a police officer, follows from the result of updating, in turn, the result of updating K_1 with the new information that Phillip P is a police officer $\vee$ the result of updating K_2 with the new information that Phillip P is a police officer. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U2, we let α represent our premises rule and we let $\beta \wedge \gamma$ represent the conjunction of our conclusion rules, β and γ. $\Pi(\alpha \wedge (\beta \wedge \gamma))$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for U8. It follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) = 20$. $\Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for U8. It follows that $\Pi(\alpha \wedge \neg(\beta \wedge \gamma)) = 17$. Further, it follows that $\Pi(\alpha \wedge (\beta \wedge \gamma)) > \Pi(\alpha \wedge \neg(\beta \wedge \gamma))$ and that a rule of the form $\alpha \rightarrow (\beta \wedge \gamma)$ is found plausible by our participants. However, the association between the premises and the conclusion is statistically non-significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners Thus, our hypothesis that human reasoning is consistent with KM postulate U8 is proven true for our participants, but not all human reasoners.

U9 has a near significant association between its premises and conclusion, according to the chi-squared test and Fisher's exact test, but a non-significant association according to the chi-squared test with Yates' correction. 67,57% (25 participants) endorsed both the premises and the conclusion. 13,51% (5 participants) committed a violation by endorsing the premises, but not the

conclusion. 10,81% (4 participants) did not endorse the premises while endorsing the conclusion. 10,81% (4 participants) endorsed neither the premises nor the conclusion. In our decomposition of postulate U9, we obtained one premises rule and one conclusion rule. The premises rule for U9 states that K is complete and the result of updating K with the new information that Mark M is a science professor $\wedge$ the new information that Mark M does enjoy solving problems is satisfiable. The conclusion rule for U9 states that the result of updating K with the new information that Mark M is a science professor, with new information that Mark M does enjoy solving problems, follows from the result of updating K with the new information that the result of updating the new information that Mark M is a science professor with the new information that Mark M does enjoy solving problems. According to the framework of possibility theory, a rule of the form $\alpha \rightarrow \beta$ is plausible iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. In the case of postulate U5, we let α represent our premises rule and we let β represent our conclusion rule. $\Pi(\alpha \wedge \beta)$ is the number of participants who endorsed both the premises and the conclusion, seen in Cell A of the contingency table for U9. It follows that $\Pi(\alpha \wedge \beta) = 25$. $\Pi(\alpha \wedge \neg\beta)$ is the number of participants who endorsed the premises, but did not endorse the conclusion, seen in Cell B of the contingency table for U9. It follows that $\Pi(\alpha \wedge \neg\beta) = 5$. Further, it follows that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ and that a rule of the form $\alpha \rightarrow \beta$ is found plausible by our participants. However, the association between the premises and conclusion is statistically non-significant to nearly significant. This means that this result is specific to our choice of material and that there is insufficient evidence to conclude that this result generalises to all English-speaking human reasoners. Thus, our hypothesis that human reasoning is consistent with the KM postulate U9 is proven true for our participants, but not all human reasoners.

5.3 Summary of Results and Discussion

In the cognitive science and psychology community, an ongoing goal is to explore the correspondence between formal logic, philosophy and human reasoning. For example, Elio and Pelletier [29] conducted experiments to determine which set of sentences to believe when additional information contradicts the initial set, a problem of belief revision. In parallel, in the AI community, an ongoing goal is to formalise models of human reasoning for use in AI-enriched systems. Ragni et al. [52, 54] have demonstrated that

classical logic fails to capture human inference, whereas non-monotonic logic has the potential to do so. In their work, they define a cognitive model as being inferentially and conceptually adequate, supported by empirical evidence. This means that, in the case of inferential adequacy, human reasoners make the same inferences as a system of formal logic. In the case of cognitive adequacy [56], the set of beliefs held by human reasoners is the same as the beliefs in the knowledge base of a system of formal logic. We have followed the inferential cognitive adequacy approach.

In previous work [5], we tested natural language (concrete) and symbolic (abstract) representations of the AGM postulates with a sample of human reasoners. We found significant ($> 50\%$) correspondence with the abstract and concrete instances of postulates R1, R3 and R4. Conversely, postulates R5 - R7 had a low correspondence while postulates R2 and R8 differed significantly in the endorsement of concrete (23,33% - 76,67%) and abstract (40% - 50%) instances. In this work, we investigated the endorsements of each component of the AGM postulates when formulated as material implication statements. In our discussion of the results for the AGM experiment, we first presented evidence for whether or not our participants found our concrete instantiations of the AGM 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 postulates, with the significance of the association between the endorsement of the premises and the endorsement of the conclusion ranging from non-significant to highly significant. The number of logical violations per postulate is generally low with a range of 0 to 4 violations ($< 12\%$ of participants). The exception is postulate R5 with 54,29% (19 participants) endorsing the premises, but not the conclusion. Upon closer inspection, the violations for R5 is similar to the percentage of participants who endorsed neither the premises nor the conclusion (48,57%) for postulate U5. We argue that this trend is explained by our participants having difficulty understanding logically equivalent sentences, with different word structures. As a result, participants reasoned cautiously and preferred to reject the conclusion. This is consistent with our understanding that the participants are not experts in formal logic and would reason cautiously by rejecting the given evidence rather than liberally asserting it. Our hypothesis that human reasoning is consistent with the AGM postulates is proven true for 3 postulates: R2, R6 and R8. For 5 postulates, our hypothesis that human reasoning is consistent with the AGM postulates is proven true for our participants, but not all human reasoners:

R1 (i), R1 (ii), R3, R4 and R7. For 1 postulate, R5, our hypothesis that human reasoning is consistent with the AGM postulates is proven false for our participants, but not all human reasoners. In the case of postulates R1 (i), R1 (ii) and R4, the general form of the postulate resembled the following rule: $\top \rightarrow \beta$. 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 3 cases, we could not conclude that this rule holds in general. We propose an additional investigation is needed to assess whether a rule of the form $\top \rightarrow \beta$ holds in general. We suggest the inclusion of tautology as a rule in a future survey. Additionally, postulates R2 and R6 share the same rule form: $\alpha \rightarrow \beta$. 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 $\alpha \rightarrow \beta$ are generally plausible. Postulates R5 and R8 also share the same rule form, given by $\alpha \rightarrow (\beta \wedge \gamma)$, where α represents a single premise and $\beta \wedge \gamma$ represents a conjunction of two conclusions. However, the results for both postulates are contradictory. Our hypothesis was proven false for R5 but proven true for R8. This indicates that rules of the form $\alpha \rightarrow (\beta \wedge \gamma)$ are both generally plausible and generally non-plausible. We argue this contradiction could have been influenced by our choice of material and participant fatigue. We propose an additional investigation is needed to assess whether a rule of the form $\alpha \rightarrow (\beta \wedge \gamma)$ holds in general with English-speaking human reasoners. A starting point for future investigation, and in parallel to the further investigation of other non-significant results for the AGM postulates, is to include shorter, concise translations of the AGM postulates in a survey.

In previous work [5], we also tested concrete and abstract representations of the KM postulates with a sample of human reasoners. We found significant ($> 50\%$) correspondence with the abstract and concrete instances of postulates U1, U3-U4 and U6. Conversely, postulates U7-U8 had a low abstract correspondence while their concrete instances were significantly endorsed. Postulates U2 and U5 were significantly endorsed in the concrete (76,7% - 90%) instances with only half of participants endorsing abstract instances. In this work, we investigated the endorsements of each component of the KM postulates when formulated as material implication statements. In our discussion of the results for the KM experiment, we first presented evidence for whether or not our participants found our concrete instantia-

tions of the KM 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 ranging from non-significant to highly significant. The number of logical violations per postulate is generally low with a range of 0 to 8 violations ($< 23\%$ of participants). The exception is postulate U8 with 48,57% (17 participants) endorsing the premises, but not the conclusion. This percentage is almost identical to the participants endorsing both the premises and conclusion (54,05%). This suggests a potential problem with the material used in the conclusion for postulate U8. Upon closer inspection, the conclusion for U8 involves two rules. We suggest that the discrepancy in the endorsement was caused by the complex formulation of the conclusion as well as the mixed (symbolic and English language) representation of the conclusion rules. Our hypothesis that human reasoning is consistent with the KM postulates is proven true for 4 postulates: U2, U3, U4 and U6. For 5 postulates, U1, U5 and U7-U9, our hypothesis that human reasoning is consistent with the KM postulates is proven true for our participants, but not all human reasoners. In the case of postulate U1, the general form of the rule is given by $\top \rightarrow \beta$. Postulates U2 and U9 share the same general form, $\alpha \rightarrow \beta \vee \gamma$, but our results for both postulates differ. While U2 and U6 apply in general, U8 applies only to our participants' reasoning. An additional investigation is needed to determine what caused the contradictory result for U8. We argue that this difference could be explained by the choice of material or our formulation of the rules. Additionally, postulates U3 and U7 share the same rule form: $\alpha \rightarrow \beta$. They too have differing results. U3 is proven true for all human reasoners while U7 is only proven true for our participants. For postulate U3, we introduced the term "satisfiable". For postulate, a new term, "complete", is introduced. 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 U7 involve complex conjunction statements which may further have confused our participants leading to a reluctance in endorsing both the premises and the conclusion. A starting point in future analyses is the inclusion of the tautological premises as a rule in the survey for participants to rank since this was an assumption not tested with our participants. However, tautological rules alone are not the cause of the range of plausibility for the KM postulates. We propose an additional

investigation is needed to assess whether all rules of the form $\top \rightarrow \beta$ and $\alpha \rightarrow \beta \wedge \gamma$. We also argue that, like the AGM postulates, the translation of the KM postulates into natural language rules when decomposed into its premises and conclusion, is not straightforward. The difference in the formulation of each postulate plays a role in whether the corresponding general rules are consistent with human reasoning.

Chapter 6

Conclusions and Future Work

In the cognitive science, psychology and AI communities, the study of human reasoning and its correspondence with formal logic is an open, ongoing scientific investigation. Our work builds on previous empirical studies involving human subjects who reason with non-monotonic information. We refined the work of Da Silva Neves et al. [47] and Benferhat et al. [9] by creating a reproducible approach for empirically investigating postulates of formal logic. This approach accounts for the effect of the premises and conclusion of each postulate and determines whether the overall postulate is found plausible with statistically significant evidence. We applied this approach to the formal theory of belief change. 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 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. 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 material from Experiment 1, which contained no bias. 30 participants were recruited from Mechanical Turk 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 linear 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 showed a modal explanation category exists for each statement and that participants were agreed on the endorsement of the premises of the statements in the survey, but differed on the endorsement of the conclusions of the statements. For Experiments 3 and 4, we recruited 50 participants on Mechanical Turk to rank the plausibility of the AGM and KM postulates on a linear scale from 1 (implausible) to 10 (extremely plausible). Each postulate is formulated in propositional logic. We decomposed the postulates into their component premises and conclusion. We then translated each component into English using the refined statements obtained from Experiment 2. We measured the association between the endorsement of the premises and the endorsement of the conclusion using the Phi-coefficient. We tested the association statistically using three measures: chi-squared, chi-squared with Yates' correction and Fisher's exact test. In our analysis, we applied possibility theory to determine whether the association between the postulates' premises and conclusion is statistically significant in general. We note that our results for the postulates in propositional logic can be lifted to higher-order logics using the appropriate representation theorems.

Our results show that our participants' reasoning tends to be consistent with the 9 AGM postulates, R1 (i), R1 (ii) and R2-R8, with the significance of the association between the endorsement of the premises and the endorsement of the conclusion ranging from non-significant to highly significant. The number of logical violations per postulate was low ($< 12\%$ of participants), except postulate R5. In our discussion, we argued that the high number of violations for postulate R5 is explained by our participants having difficulty understanding logically equivalent sentences, with different word structures. Our hypothesis that human reasoning is consistent with the AGM postulates is proven true for 3 postulates: R2, R6 and R8. For 5 postulates, our hypothesis that human reasoning is consistent with the AGM postulates is proven true for our participants, but not all human reasoners: R1 (i), R1 (ii), R3, R4 and R7. For 1 postulate, R5, our hypothesis that human reasoning is consistent with the AGM postulates is proven false for our participants, but not all human reasoners. We provided evidence that our participants found rules

of the form $\alpha \rightarrow \beta$ plausible in general. Further, we obtained contradictory findings in our analysis of postulates R5 and R8. Despite sharing the same rule form, our hypothesis was proven false for R5 but proven true for R8. Their rule form is given by $\alpha \rightarrow \beta \wedge \gamma$, where α represents a single premises and $\beta \wedge \gamma$ represents a conjunction of two conclusions. In our discussion, we argued that this contradiction could have been influenced by our choice of material and participant fatigue.

Our results also show that our participants' reasoning tends to be consistent with the 9 KM postulates, U1 - U9, with the significance of the association between the endorsement of the premises and the endorsement of the conclusion ranging from non-significant to highly significant. The number of logical violations per postulate is generally low ($< 23\%$ of participants), except postulate U8. In our discussion, we argued that the high number of violations for postulate U8 was caused by the complex formulation of the conclusion as well as the mixed (symbolic and English language) representation of the conclusion rules. Our hypothesis that human reasoning is consistent with the KM postulates is proven true for 4 postulates: U2, U3, U4 and U6. For 5 postulates, U1, U5 and U7-U9, our hypothesis that human reasoning is consistent with the KM postulates is proven true for our participants, but not all human reasoners. In the case of postulate U1, the general form of the rule is given by $\top \rightarrow \beta$. Postulates U2 and U9 share the same general form, $\alpha \rightarrow \beta \vee \gamma$, but our results for both postulates differ. While U2 and U6 apply in general, U8 applies only to our participants' reasoning. An additional investigation is needed to determine what caused the contradictory result for U8. In our discussion, we argued that this difference could be explained by the choice of material or our formulation of the rules. Additionally, postulates U3 and U7 share the same rule form: $\alpha \rightarrow \beta$. They too have differing results. U3 is proven true for all human reasoners while U7 is only proven true for our participants. For postulate U3, we introduced the term "satisfiable". For postulates U7 and U9, a new term, "complete", was introduced. We suggest that our participants wrestled with the meaning of complete belief sets. This was not unexpected as our participants were not experts in formal logic. Further, the concrete rules used in U7 involved complex conjunction statements which may further have confused our participants, leading to a reluctance in endorsing both the premises and the conclusion.

We propose an additional investigation of rules of the form $\alpha \rightarrow \beta \wedge \gamma$ to determine whether it holds in general with English-speaking human reasoners. The investigation of tautological rules of the form $\top \rightarrow \beta$ is also

necessary to validate our assumption that it is found plausible by human reasoners. We submit that the translation of the AGM and KM postulates into English rules is not straightforward. The difference in the formulation of each postulate plays a significant role in whether the corresponding English rules are consistent with human reasoning. Our immediate resolve is to review our data design and data analysis steps. We aim to include non-formal, concise translations of the postulates in a future survey. An avenue for further investigation is the use of non-parametric statistical methods to evaluate the significance of the correspondence between human reasoning and postulates of belief change. This will complement our investigation that used parametric statistical methods for data analysis, and enable a comparison of the efficacy of each method.

Bibliography

[1] Alchourrón, C. E., Gärdenfors, P., and Makinson, D. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic 50* (1985), 510–530. 9, 15, 17

[2] Andreotta, M., Nugroho, R., Hurlstone, M., Boschetti, F., Farrel, S., Walker, I., and Paris, C. Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis. *Behavior research methods 51*, 4 (2019), 1766–1781. 23

[3] Antoy, S., Hanus, M., and Teegen, F. Synthesizing set functions. In *International Workshop on Functional and Constraint Logic Programming* (2018), Springer, pp. 93–111. 19

[4] Arechar, A., Kraft-Todd, G., and Rand, D. Turking overtime: how participant characteristics and behavior vary over time and day on amazon mechanical turk. *Journal of the Economic Science Association 3*, 1 (2017), 1–11. 32

[5] Baker, C., Denny, C., Freund, P., and 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)* (2020), CCIS, Springer, pp. 119–219. 9, 10, 32, 90, 91

[6] Baker, C., and Meyer, T. Belief change in human reasoning: An empirical investigation on mturk. In *Proceedings of the Second Southern African Conference for Artificial Intelligence Research (SACAIR 2021)* (2021), p. To appear. 10

[7] Batchelder, W. H., and Riefer, D. M. Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review 6*, 1 (1999), 57–86. 9

[8] Ben-Ari, M. Propositional logic: Formulas, models, tableaux. In *Mathematical Logic for Computer Science*. Springer, 2012, pp. 7–47. 11

[9] Benferhat, S., Bonnefon, J. F., and da Silva Neves, R. An overview of possibilistic handling of default reasoning, with experimental studies. *Synthese 146*, 1 (2005), 53–70. 8, 94

[10] Benferhat, S., Dubois, D., and Prade, H. Representing default rules in possibilistic logic. In *Proceedings of 3rd International conference on principles of knowledge representation and reasoning (KRâĂŹ92)* (1992), pp. 673–684. 76

[11] Benferhat, S., Dubois, D., and Prade, H. Nonmonotonic reasoning, conditional objects and possibility theory. *Artificial Intelligence 92*, 1-2 (1997), 259–276. 20

[12] Benferhat, S., Dubois, D., and Prade, H. Practical handling of exception-tainted rules and independence information in possibilistic logic. *Applied intelligence 9*, 2 (1998), 101–127. 20

[13] Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. Redefine statistical significance. *Nature human behaviour 2*, 1 (2018), 6–10. 29

[14] Bentley, F., Daskalova, N., and White, B. Comparing the reliability of amazon mechanical turk and survey monkey to traditional market research surveys. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017), pp. 1092–1099. 31

[15] Bergmann, M., Moor, J., and Nelson, J. *The logic book*, vol. 3. McGraw-Hill, 1998. 11

[16] Berinsky, A., Huber, G., and Lenz, G. Evaluating online labor markets for experimental research: Amazon.comâĂŹs mechanical turk. *Political Analysis 20*, 3 (2012), 351–368. 32

[17] Brachman, R., Levesque, H., and Pagnucco, M. *Knowledge Representation and Reasoning*. The Morgan Kaufmann Series in Artificial Intelligence. Elsevier Science & Technology, San Francisco, 2004. 7

[18] Byrne, R. M. Suppressing valid inferences with conditionals. *Cognition 31*, 1 (1989), 61–83. 8

[19] Coppock, A. Generalizing from survey experiments conducted on mechanical turk: A replication approach. *Political Science Research and Methods 7*, 3 (2019), 613–628. 32

[20] Cramer, H. Mathematical methods of statistics, princeton univ. *Press, Princeton, NJ* (1946). 27

[21] Creignou, N., Ktari, R., and Papini, O. Belief update within propositional fragments. *Journal of Artificial Intelligence Research 61* (2018), 807–834. 17

[22] Creswell, J. W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 4th ed. SAGE Publications, Inc., January 2014. 23

[23] Derrick, B., and White, P. Comparing two samples from an individual likert question. *International Journal of Mathematics and Statistics 18*, 3 (2017). 9

[24] Dubois, D., and Prade, H. Non-monotonic reasoning and uncertainty theories. *Nonmonotonic Reasoning. Essays Celebrating its 30th Anniversary.. Gerhard Brewka, Victor Marek, Miroslaw Truszczynski (Eds.), College Publications* (2011), 141–176. 20

[25] Dubois, D., and Prade, H. Possibilistic logic - an overview. In *Handbook of the History or Logic* (2014), vol. 9, pp. 197–255. 19, 20

[26] Dubois, D., and Prade, H. A crash course on generalized possibilistic logic. In *International Conference on Scalable Uncertainty Management* (2018), Springer, pp. 3–17. 19

[27] Dubois, D., and Prade, H. Possibility theory and possibilistic logic: Tools for reasoning under and about incomplete information. In *4th IFIP International Conference on Intelligence Science-TC 12: Artificial Intelligence (ICIS 2021)* (2021), vol. 623, Springer, pp. 79–89. 19

[28] Edlund, J., Sagarin, B., Skowronski, J., Johnson, S., and Kutter, K. Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk. *Personality and Social Psychology Bulletin 35*, 5 (2009), 635–642. 32

[29] Elio, R., and Pelletier, F. Belief change as propositional update. *Cognitive Science 21*, 4 (1997), 419–460. 89

[30] Fermé, E., and Hansson, S. O. *Belief Change: Introduction and Overview*. SpringerBriefs in Intelligent Systems. Springer International Publishing AG, Cham, 2018. 14

[31] Fisher, R. A. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society 85*, 1 (1922), 87–94. 29

[32] Gärdenfors, P. Epistemic importance and minimal changes of belief. *Australasian Journal of Philosophy 62*, 2 (1984), 136–157. 17

[33] Gärdenfors, P., and Makinson, D. Revisions of knowledge systems using epistemic entrenchment. In *Proceedings of the 2nd conference on Theoretical aspects of reasoning about knowledge* (1988), pp. 83–95. 15

[34] Grootswagers, T. A primer on running human behavioural experiments online. *Behavior research methods* (2020), 1–4. 31, 33

[35] Hentschke, H., and Stüttgen, M. Computation of measures of effect size for neuroscience data sets. *European Journal of Neuroscience 34*, 12 (2011), 1887–1894. 29, 53

[36] Herzig, A., Lang, J., and Marquis, P. Propositional update operators based on formula/literal dependence. *ACM Transactions on Computational Logic 14*, 3 (2013), 1–31. 17

[37] Herzig, A., and Rifi, O. Propositional belief base update and minimal change. *Artificial Intelligence 115*, 1 (1999), 107–138. 17

[38] Johnson, V., Payne, R., Wang, T., Asher, A., and Mandal, S. On the reproducibility of psychological science. *Journal of the American Statistical Association 112*, 517 (2017), 1–10. 29, 53

[39] Katsuno, H., and Mendelzon, A. A unified view of propositional knowledge base updates. *Artificial Intelligence 11*, 2 (1989), 1413–1419. 17

[40] Katsuno, H., and Mendelzon, A. Propositional knowledge base revision and minimal change. *Artificial Intelligence 3*, 52 (1991), 263–294. 16

[41] Katsuno, H., and Mendelzon, A. O. On the difference between updating a knowledge base and revising it. *Belief revision* (1991), 183. 9, 17, 19

[42] Kraus, S., Lehmann, D., and Magidor, M. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence 44* (1990), 167–207. 8, 20

[43] Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., et al. Justify your alpha. *Nature Human Behaviour 2*, 3 (2018), 168–171. 29

[44] Machleid, F., Kaczmarczyk, R., Johann, D., Balčiūnas, J., Atienza-Carbonell, B., von Maltzahn, F., and Mosch, L. Perceptions of digital health education among european medical students: Mixed methods survey. *Journal of medical Internet research. 22*, 8 (2020-8-14). 23

[45] Miller, T.and Muise, C. Belief update for proper epistemic knowledge bases. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA, 2016), IJCAI'16, AAAI Press, pp. 1209–1215. 17

[46] Nadarzynski, T., Miles, O., Cowie, A., and Ridge, D. Acceptability of artificial intelligence (ai)-led chatbot services in healthcare: A mixed-methods study. *Digital health 5* (2019), 2055207619871808. 23

[47] Neves, R., Bonnefon, J., and Raufaste, E. An empirical test of patterns for nonmonotonic inference. *Annals of Mathematics and Artificial Intelligence 34*, 1-3 (2002), 107–130. 8, 94

[48] O'Halloran, K., Tan, S., Pham, D., Bateman, J., and Vande Moere, A. A digital mixed methods research design: Integrating multimodal analysis with data mining and information visualization for big data analytics. *Journal of Mixed Methods Research 12*, 1 (2018), 11–30. 23

[49] Oppenheimer, D., Meyvis, T., and Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology 45* (2009), 867–872. 32

[50] Orne, M. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist 17*, 11 (1962), 776. 32

[51] Paolacci, G., and Chandler, J. Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science 25* (2014), 184–188. 33

[52] Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., and Tse, A. Formal nonmonotonic theories and properties of human defeasible reasoning. *Minds and Machines 27* (2017), 79–117. 8, 89

[53] Ragni, M., Eichhorn, C., and Kern-Isberner, G. Simulating human inferences in light of new information: A formal analysis. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 16)* (2016), S. Kambhampati, Ed., IJCAI Press, pp. 2604–2610. 8

[54] Ragni, M., Kern-Isberner, G., Beierle, C., and Sauerwald, K. Cognitive logics–features, formalisms, and challenges. In *ECAI 2020*. IOS Press, 2020, pp. 2931–2932. 8, 89

[55] Rana, R., and Singhal, R. Chi-square test and its application in hypothesis testing. *Journal of the Practice of Cardiovascular Sciences 1*, 1 (2015), 69. 29, 53

[56] Renz, J., Rauh, R., and Knauff, M. Towards cognitive adequacy of topological spatial relations. In *Spatial Cognition II*. Springer, 2000, pp. 184–197. 90

[57] Ribeiro, J., Nayak, A., and Wassermann, R. Belief update without compactness in non-finitary languages. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (2019), pp. 1858–1864. 17

[58] Riessman, C. K. *Narrative methods for the human sciences*. Sage, 2008. 22

[59] Russell, S., and Norvig, P. *Artificial Intelligence: A Modern Approach*, 3 ed. Prentice Hall, 2010. 7

[60] Schon, C., Siebert, S., and Stolzenburg, F. The corg project: Cognitive reasoning. *KI-Künstliche Intelligenz 33*, 3 (2019), 293–299. 8

[61] Souza, E., and Negri, T. First prospects in a new approach for structure monitoring from gps multipath effect and wavelet spectrum. *Advances in Space Research 59*, 10 (2017), 2536–2547. 28

[62] Springer, V., Martini, P., Lindsey, S., and Vezich, I. Practice-based considerations for using multi-stage survey design to reach special populations on amazonâĂŹs mechanical turk. *Survey Practice 9*, 5 (2016), 1–8. 32, 33

[63] Van Harmelen, F., Lifschitz, V., and Porter, B. *Handbook of knowledge representation*. Elsevier, 2008. 7

[64] Weijters, B., Millet, K., and Cabooter, E. Extremity in horizontal and vertical likert scale format responses. some evidence on how visual distance between response categories influences extreme responding. *International Journal of Research in Marketing 38*, 1 (2021), 85–103. 9

[65] Winslett, M. Reasoning about action using a possible models approach. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI 88)* (1988), AAAI Press, pp. 89–93. 17

[66] Yates, F. Contingency tables involving small numbers and the $\mathcal{X}^2$ test. *Supplement to the Journal of the Royal Statistical Society 1*, 2 (1934), 217–235. 29

[67] Zadeh, L. A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems 1*, 1 (1978), 3–28. 19

[68] Zhao, F., Xu, J., and Lin, Y. Similarity measure for patients via a siamese cnn network. In *Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing* (November 2018), Springer, pp. 319–328. 28

Appendix A

Survey URLs

For convenience, the URLs to our surveys are included in Table A.1.

Table A.1: 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

Appendix B

Survey Material

B.1 Experiment 1

In Figures B.1–B.3, we show the survey material for experiment 1.

B.2 Experiment 2

In Figures B.4–B.6, we show the survey material for experiment 2.

B.3 Experiment 3

In Figures B.7–B.10, we show the survey material for experiment 3.

B.4 Experiment 4

In Figures B.11–B.15, we show the survey material for experiment 4.

Survey material: pre-experiment

Statement 1

If Max S is a doctor then Max S does have an illegible handwriting

Statement 2

If Hilda P is a school teacher then Hilda P does love children

Statement 3

If Kevin A is a police officer then Kevin A does eat doughnuts

Statement 4

If Amy W is a lawyer then Amy W does wear a suit to work

Statement 5

If Mark M is a science professor then Mark M does enjoy solving problems

Statement 6

If Zeeta M is classical pianist then Zeeta M is a classical pianist

Statement 7

If Patsy V is a politician then Patsy V does have ethical conduct

Statement 8

If Luke B is a programmer then Luke B does drink lots of coffee

Statement 9

If Cynthia K is a professional cyclist then Cynthia K does cycle to work

Statement 10

If Harold C is an office manager then Harold C does have a short temper

Statement 11

If Margaret R is a baker then Margaret R does get up early in the morning to bake bread

Figure B.1: Survey material for experiment 1 (1 of 3)

Statement 12

If Eric V is a football player from Slovenia then Eric V is an football player from Slovenia

Statement 13

If Eddie T is a school principal then Eddie T does make long speeches

Statement 14

If Anita J is a female then Anita J is a feminist

Statement 15

If Carl S is a millionaire then Carl S does play golf

Statement 16

If Wilma D is a car owner then Wilma D does pay insurance

Statement 17

If Jeremy N is a detective then Jeremy N does smoke cigarettes

Statement 18

If Constance Y works in an office with two windows then Constance Y works in an office with two windows

Statement 19

If Nicole A is a cancer patient then Nicole A is terminally ill

Statement 20

If Chris P is a waiter then Chris P has profound knowledge of the menu

Statement 21

If Daisy M is a crying baby then Daisy M is hungry

Statement 22

If Tom F is a psychopath then Tom F does have a twitchy eye

Figure B.2: Survey material for experiment 1 (2 of 3)

Statement 23

If Kelsey H is a hairdresser then Kelsey H does have colour-treated hair

---

Statement 24

If Jessica B is a yoga instructor then Jessica B is a yoga instructor

---

Statement 25

If Daniel U is a circus clown then Daniel U does have large feet

---

Statement 26

If Penny K wears spectacles then Penny K is near-sighted

Statement 27

If Quentin O is a nurse then Quentin O does have a caring bedside manner

---

Statement 28

If Isabelle C is a test invigilator then Isabelle C does carry a spare set of pens

---

Statement 29

If Rory Z is an accountant then Rory Z is a mathematics boffin

---

Statement 30

If Noel W is a strong firefighter, then Noel W is a strong firefighter

---

Ends.

Figure B.3: Survey material for experiment 1 (3 of 3)

Survey material: general reasoning

Statement 1

If Damian S is an aeroplane pilot then Damian S has visited different countries

Statement 2

If Rochelle P is a ballerina then Rochelle P does listen to classical music

Statement 3

If Jacob B is a truck driver then Jacob B does drive at night

Statement 4

If Amy W is a lawyer then Amy W does wear a suit to work

Statement 5

If Mark M is a science professor then Mark M does enjoy solving problems

Statement 6

If Zeeta M is classical pianist then Zeeta M is a classical pianist

Statement 7

If Michaela W is an actress then Michaela W is hard to please

Statement 8

If Liam F is a computer science student then Liam F is a hacker

Statement 9

If Carina K is a lawyer then Carina K does tell lies

Statement 10

If Harold C is an office manager then Harold C does have a short temper

Statement 11

If Margaret R is a baker then Margaret R does get up early in the morning to bake bread

Figure B.4: Survey material for experiment 2 (1 of 3)

Statement 12

If Eric V is a football player from Slovenia then Eric V is an football player from Slovenia

Statement 13

If Eddie T is a school principal then Eddie T does make long speeches

Statement 14

If Anita J is a female then Anita J is a feminist

Statement 15

If John T is a farmer then John T does cultivate crops

Statement 16

If Wilma D is a car owner then Wilma D does pay insurance

Statement 17

If Phillip P is a police officer then Phillip P can arrest a criminal

Statement 18

If Constance Y works in an office with two windows then Constance Y works in an office with two windows

Statement 19

If Nicole A is a cancer patient then Nicole A is terminally ill

Statement 20

If Chris P is a waiter then Chris P has profound knowledge of the menu

Statement 21

If Daisy M is a crying baby then Daisy M is hungry

Statement 22

If Bryce V is a dentist then Bryce V does have perfect teeth

Figure B.5: Survey material for experiment 2 (2 of 3)

Statement 23

If Alicia E watches movies then Alicia E does watch romantic comedies

---

Statement 24

If Jessica B is a yoga instructor then Jessica B is a yoga instructor

---

Statement 25

If Steven R keeps a pet dog then Steven R does feed his dog every day

---

Statement 26

If Penny K wears spectacles then Penny K is near-sighted

Statement 27

If Quentin O is a nurse then Quentin O does have a caring bedside manner

---

Statement 28

If Isabelle C is a test invigilator then Isabelle C does carry a spare set of pens

---

Statement 29

If Rory Z is an accountant then Rory Z is a mathematics boffin

---

Statement 30

If Noel W is a strong firefighter, then Noel W is a strong firefighter

---

Ends.

Figure B.6: Survey material for experiment 2 (3 of 3)

Survey material: AGM belief revision

R1 conclusion 1 rule

K includes not only the beliefs contained in it, but also the consequences which follow from K.

R1 conclusion 2 rule

Given

- Zeeta M is a classical pianist (new information)

The result of revising K with the new information that Zeeta M is a classical pianist, also contains the information which follows from the result of revising K with the new information that Zeeta M is a classical pianist.

R2 premise rule

Given

- Jacob B is a truck driver (new information)

Jacob B does drive at night follows from revising K with the the new information that Jacob B is a truck driver.

R2 conclusion rule

Given

- Jacob B is a truck driver (new information)

Jacob B does drive at night follows from expanding K with the new information that Jacob B is a truck driver.

R3 premise rule

Given

- Jessica B is a yoga instructor (new information)

K is satisfiable with respect to the new information that Jessica B is a yoga instructor.

Figure B.7: Survey material for experiment 3 (1 of 4)

R3 conclusion 1 rule

Given
- Jessica B is a yoga instructor (new information)

Jessica B does teach breathing exercises follows from the expansion of K with the new information that Jessica B is a yoga instructor.

R3 conclusion 2 rule

Given
- Jessica B is a yoga instructor (new information)

Jessica B does teach breathing exercises follows from the revision of K with the new information that Jessica B is a yoga instructor.

R4 conclusion rule

Given
- Chris P is a waiter (new information)

The new information that Chris P is waiter is contained in the revision of K with the new information that Chris P is a waiter.

R5 premise rule

Given
- If Noel W is a firefighter then Noel W is strong (new information)
- Either Noel W is not a firefighter or Noel W is strong (new information)

The new information that if Noel W is a firefighter then Noel W is strong is equivalent to the new information that either Noel W is not a firefighter or Noel W is strong.

R5 conclusion 1 rule

Given
- If Noel W is a firefighter then Noel W is strong (new information)
- Either Noel W is not a firefighter or Noel W is strong (new information)

Noel W does save lives follows from the revision of K with the new information that Noel W is a strong firefighter.

Figure B.8: Survey material for experiment 3 (2 of 4)

R5 conclusion 2 rule

Given
- If Noel W is a firefighter then Noel W is strong (new information)
- Either Noel W is not a firefighter or Noel W is strong (new information)

Noel W does save lives follows from the revision of K with the new information that either Noel W is not a firefighter or Noel W is strong.

R6 premise rule

Given
- Wilma D is a car owner (new information)

The new information that Wilma D is a car owner, is consistent.

R6 conclusion rule

Given
- Wilma D is a car owner (new information)

The revision of K with the new information that Wilma D is a car owner, is consistent.

R7 premise rule

Given
- Phillip P is a police officer and Phillip P can arrest a criminal (new information)

Phillip P does carry a gun follows from the revision of K with the new information that Phillip P is a police officer and Phillip P can arrest a criminal.

R7 conclusion rule

Given
- Philip P is a police officer (new information)
- Philip P can arrest a criminal (new information)

Phillip P does carry a gun follows from the result of first revising K with the new information that Philip P is a police officer and then expanding with the new information that Philip P can arrest a criminal.

Figure B.9: Survey material for experiment 3 (3 of 4)

R8 premise rule

Given
- Mark M is a science professor (new information)
- Mark M does enjoy solving problems (new information)

The revision of K with the new information that Mark M is a science professor is satisfiable with respect to the new information that Mark M does enjoy solving problems.

R8 conclusion 1 rule

Given
- Mark M is a science professor (new information)
- Mark M does enjoy solving problems (new information)

Mark M is a good teacher follows from the result of first revising K with the new information that Mark M is a science professor and then expanding with the new information that Mark M does enjoy solving problems.

R8 conclusion 2 rule

Given
- Mark M is a science professor and Mark M does enjoy solving problems (new information)

Mark M is a good teacher follows from the result of revising K with the new information that Mark M is a science professor and Mark M does enjoy solving problems.

Ends.

Figure B.10: Survey material for experiment 3 (4 of 4)

Survey material: KM belief update

U1 conclusion rule

Given
1. Jacob B is a truck driver (new information)

The new information (1) follows from the result of updating ψ with the new information (1).

U2 premise rule

Given
1. Noel W is a strong firefighter (new information)

The new information (1) follows from ψ.

U2 conclusion 1 rule

Given
1. Noel W is a strong firefighter (new information)

ψ follows from the result of updating ψ with the new information (1).

U2 conclusion 2 rule

Given
1. Noel W is a strong firefighter (new information)

The result of updating ψ with the new information (1), follows from ψ.

U3 premise rule

Given
1. Eric V is a football player from Slovenia (new information)

Both ψ and the new information (1), is satisfiable.

U3 conclusion rule

Given
1. Eric V is a football player from Slovenia (new information)

The result of updating ψ with the new information (1), is satisfiable.

Figure B.11: Survey material for experiment 4 (1 of 5)

U4 premise 1 rule

Given
1. Zeeta M is a classical pianist (new information)
2. Zeeta M is a classical pianist (new information)
3. ψ1 (a belief base)
4. ψ2 (a belief base, different to ψ1)

ψ2 follows from ψ1.

U4 premise 2 rule

Given
1. Zeeta M is a classical pianist (new information)
2. Zeeta M is a classical pianist (new information)
3. ψ1 (a belief base)
4. ψ2 (a belief base, different to ψ1)

ψ1 follows from ψ2.

U4 premise 3 rule

Given
1. Zeeta M is a classical pianist (new information)
2. Zeeta M is a classical pianist (new information)
3. ψ1 (a belief base)
4. ψ2 (a belief base, different to ψ1)

The new information (1) follows from the new information (2).

U4 premise 4 rule

Given
1. Zeeta M is a classical pianist (new information)
2. Zeeta M is a classical pianist (new information)
3. ψ1 (a belief base)
4. ψ2 (a belief base, different to ψ1)

The new information (2) follows from the new information (1).

Figure B.12: Survey material for experiment 4 (2 of 5)

U4 conclusion 1 rule

Given

1. Zeeta M is a classical pianist (new information)
2. Zeeta M is a classical pianist (new information)
3. $\psi 1$ (a belief base)
4. $\psi 2$ (a belief base, different to $\psi 1$)

The result of updating $\psi 2$ with the new information (2), follows from the result of updating $\psi 1$ with the new information (1).

U4 conclusion 2 rule

Given

1. Zeeta M is a classical pianist (new information)
2. Zeeta M is a classical pianist (new information)
3. $\psi 1$ (a belief base)
4. $\psi 2$ (a belief base, different to $\psi 1$)

The result of updating $\psi 1$ with the new information (1), follows from the result of updating $\psi 2$ with the new information (2).

U5 conclusion rule

Given

1. Wilma D is a car owner (new information)
2. Wilma D does pay insurance (new information)
3. The result of updating ψ with the information that Wilma D is a car owner $\wedge$ Wilma D does pay insurance (statement)
4. Wilma D is a car owner $\wedge$ Wilma D does pay insurance (statement)

The result of updating ψ with statement (4), follows from statement (3).

U6 premise rule

Given

1. Chris P is a waiter (new information)
2. Chris P has profound knowledge of the menu (new information)

The new information (2) follows from the result of updating ψ with new information (1), and the new information (1) follows from the result of updating ψ with new information (2).

Figure B.13: Survey material for experiment 4(3 of 5)

U6 conclusion 1 rule

Given
1. Chris P is a waiter (new information)
2. Chris P has profound knowledge of the menu (new information)

The result of updating ψ with the new information (2) follows from the result of updating ψ with the new information (1).

U6 conclusion 2 rule

Given
1. Chris P is a waiter (new information)
2. Chris P has profound knowledge of the menu (new information)

The result of updating ψ with the new information (1) follows from the result of updating ψ with the new information (2).

U7 premise rule

ψ is complete.

U7 conclusion rule

Given
1. Jessica B is a yoga instructor (new information)
2. Jessica B does teach breathing exercises (new information)
3. new information (1) $\wedge$ new information (2) (statement)
4. The result of updating ψ with the new information (1) $\wedge$ the result of updating ψ with the new information (2) (statement)

The result of updating ψ with statement (3), follows from statement (4).

U8 conclusion 1 rule

Given
1. Philip P is a police officer (new information)
2. $\psi 1$ (a belief base)
3. $\psi 2$ (a belief base, different to $\psi 1$)
4. $\psi 1 \vee \psi 2$ (statement)
5. The result of updating $\psi 1$ with the new information (1) (statement)
6. The result of updating $\psi 2$ with the new information (1) (statement)
7. statement (5) $\vee$ statement (6) (statement)

Statement (7) follows from the result of updating statement (4) with the new information (1).

Figure B.14: Survey material for experiment 4 (4 of 5)

U8 conclusion 2 rule

Given
1. Philip P is a police officer (new information)
2. $\psi 1$ (a belief base)
3. $\psi 2$ (a belief base, different to $\psi 1$)
4. $\psi 1 \vee \psi 2$ (statement)
5. The result of updating $\psi 1$ with the new information (1) (statement)
6. The result of updating $\psi 2$ with the new information (1) (statement)

statement (5) ∨ statement (6) (statement)

The result of updating statement (4) with the new information (1), follows from statement (7).

U9 premise rule

Given
1. Mark M is a science professor (new information)
2. Mark M does enjoy solving problems (new information)
3. new information (1) ∧ new information (2) (statement)
4. The result of updating ψ with the new information (1) (statement)
5. The result of updating the new information (1) with new information (2) (statement)

ψ is complete, and the result of statement (4) ∧ new information (2) is satisfiable.

U9 conclusion rule

Given
1. Mark M is a science professor (new information)
2. Mark M does enjoy solving problems (new information)
3. new information (1) ∧ new information (2) (statement)
4. The result of updating ψ with the new information (1) (statement)
5. The result of updating the new information (1) with new information (2) (statement)

The result of updating statement (4) with new information (2), follows from the result of updating ψ with statement (5).

Ends.

Figure B.15: Survey material for experiment 4 (5 of 5)

Appendix C

Supplementary Information

The Github repository for this work, containing supplementary material and code scripts, can be accessed via this URL, `https://tinyurl.com/2p98m76n`.