UNIVERSITY OF CAPE TOWN
DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours
Final Paper 2019

Title: On the relation between defeasible and human reasoning

Author: Clayton Baker

Project Abbreviation: CDR

Supervisor(s): Prof. Tommie Meyer

Material related to this project, namely the survey questions on Google Forms, can be found by following this link: https://forms.gle/9yggoXF9U6uhxqCm6

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	
Theoretical Analysis	*0*	*25*	5
Experiment Design and Execution	*0*	*20*	20
System Development and Implementation	*0*	*20*	
Results, Findings and Conclusion	*10*	*20*	20
Aim Formulation and Background Work	*10*	*15*	15
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	
Total marks		**80**	

On the relation between defeasible and human reasoning

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

ABSTRACT

Humans reason in different ways. In the AI community, computational models of reasoning have been developed to represent human reasoning. Initially, classical models were introduced in which reasoning systems depended on a fixed knowledge base and a fixed set of conditions. However, the diversity of human reasoning is not well represented using the classical reasoning model. Instead, when humans receive conflicting information, a more flexible reasoning approach is required. This is true for AI systems as well. Forms of defeasible reasoning in the AI community are closely related to how humans reason. We focus on the properties of defeasible reasoning, proposed by Kraus-Lehmann-Magidor (KLM) [17], and the extent to which these apply to human reasoning. We reintroduce the notion of non-monotonic reasoning and provide related work for defeasible reasoning and related work for two similar models of non-monotonic reasoning. We document the design and execution of a study that we have performed to test empirically the extent to which each of the KLM [17] properties are appropriate, when presented to human reasoners in the form of natural and abstract reasoning scenarios. We have evidence that the KLM [17] property of **Or** and the defeasible reasoning property of **Transitivity** have a strong positive agreement with human reasoning. We provide evidence in support of a normative relationship between defeasible and human reasoning. We also discuss the results of this study, as well as further insight into the relationship between defeasible and human reasoning. Studies determining the extent to which belief revision and belief update respectively compare with human reasoning, are currently being undertaken by other AI researchers. The comparison of our results with the results from studies determining the correspondence between human reasoning and different forms of non-monotonic reasoning, such as belief revision and belief update, is left for future work.

CCS CONCEPTS

• **Theory of computation** → **Logic**; • **Computing methodologies** → **non-monotonic reasoning, defeasible reasoning**; *Typicality-based reasoning and retractable inferences*;

KEYWORDS

propositional logic, defeasible reasoning, non-monotonic reasoning, human reasoning

1 INTRODUCTION

Humans are confronted with many situations every day in which they are required to make a decision. Humans are also known to reason in diverse ways. In the AI community, different reasoning models which represent human reasoning exist as well. Initially, classical reasoning was introduced as a fixed-parameter reasoning paradigm [8, 10, 11, 13, 32]. Classical reasoning, however, is insufficient to represent human reasoning because it does not have the appropriate mechanisms to deal with atypical reasoning situations. On the other hand, non-monotonic reasoning has been introduced as a model that accounts for changes in a conclusion that a reasoner draws when additional information is presented to the reasoner. Non-monotonic reasoning is characterised by conditional statements contained in a knowledge base. Each conditional has two identifiable parts. The first part is the antecedent, which is the set of given information that is available to the reasoner. The remaining part of the conditional is the consequent, which is validated when then antecedent is warranted. To illustrate this, we will now describe an example. Consider the following conditional statement: "if Lisa has a cake to bake, then she will use an oven". The antecedent in this example is "if Lisa has a cake to bake". If this condition is met, in other words, if Lisa indeed has a cake to bake, then the reasoner can reasonably infer the consequent "she will use an oven". Our interest lies with the term "reasonably". This paper aims to identify the circumstances under which human reasoning corresponds to a form of non-monotonic reasoning called defeasible reasoning.

In our study, we have considered the six primary defeasible reasoning properties proposed by Kraus-Lehmann-Magidor (KLM) [17]. We have conducted a survey posing ordinary, hypothetical and abstract reasoning scenarios to human participants, to test our hypothesis empirically. Some of the anticipated outcomes of our study include determining which KLM [17] properties are true with human reasoning and which are not. We also expect from our study that human reasoning has a normative relationship when compared to defeasible reasoning. We then present and discuss our results. The rest of the paper is made up as follows. In Section 2 we will describe how non-monotonic reasoning relates to human reasoning. In Section 2 we will also describe related work for defeasible reasoning, belief revision and belief update as well as define a language for classical reasoning using propositional logic. In Section 3, we will present a formal representation of defeasible reasoning logic, and describe the defeasible reasoning properties pertinent to our study. In Section 4, we outline our research questions, describe our anticipated outcomes and describe the expected impact of our research. In Section 5, we describe the design and implementation of our study as well as our testing ad evaluation strategy. In Section 5, we also provide the proposed theoretical contributions of this paper. In Section 6, we present and discuss the outcomes of our study on Mechanical Turk. In Section 7, we provide the empirical results from our study and answer some of our research questions. Lastly, we emphasise some of the important conclusions from our study in Section 8.

2 BACKGROUND

The representations of natural reasoning in the AI community have been adapted to reflect the diversity in human reasoning. With all reasoning, a common thread exists: there is an amount of knowledge known to the reasoner and the reasoner uses this knowledge in some way to make an inference. Initially, AI systems were programmed with classical reasoning. In classical reasoning, a conclusion made by an agent, or reasoner, is drawn from a fixed set of information. Classical reasoning is insufficient to represent human reasoning because it does not have the appropriate mechanisms to deal with atypical reasoning situations. On the other hand, non-monotonic reasoning has been introduced as a model that deals with atypical reasoning situations to some extent. Models of non-monotonic reasoning account for changes in an inference made by a reasoner when additional information is presented to the reasoner. In the following example, the reasoner is also a vehicle driver.

Example. Consider the following scenario:
(i) a driver is waiting at a traffic light
(ii) the traffic light has turned green

Can the driver infer that it is safe to continue driving? In ordinary traffic circumstances, this inference is plausible. Should it be made known to the driver that, at the moment the traffic light turned green, a pedestrian was crossing the road in front of the driver, then the driver (with some degree of compassion) would infer that is not safe to continue driving and that he should wait for the pedestrian to cross over before driving off.

Pelletier et al. [25] discussed non-monotonic reasoning and the influence of psychologism in human reasoning. Lehmann [19] presented stereotypical reasoning as a form of natural non-monotonic reasoning. Gärdenfors et al. [12] described the notion of expectations in the context of non-monotonic inferences. More specifically, there are several forms of non-monotonic reasoning that have been widely studied. They include defeasible reasoning [17, 20, 26, 27], belief revision [9, 15, 29] and belief update [14, 16]. We consider the formal model of defeasible reasoning and how it relates to human reasoning. In the following section, we first define a language for classical reasoning that will be based on conventions from propositional logic. This language will further be used throughout the paper and forms the basis of the defeasible reasoning language we use as well. Lastly, in this section, we describe related work for defeasible reasoning, belief revision and belief update.

2.1 Propositional Logic

In this section, we formally describe the language for classical reasoning. The language we shall use, L, is based on propositional logic. All the different beliefs of the world will be formulated in L. We will use Greek letters such as α, β and γ etc. as variables over sentences in L. It will be assumed that L is closed under applications of the boolean connectives $\rightarrow$ (implication), $\wedge$ (conjunction), $\vee$ (disjunction) and $\neg$ (negation). These connectives can be applied recursively to statements formulated in L. We will represent a knowledge base with notation $\mathcal{K}$. The knowledge base contains all the statements that are both known and believed to be true by the reasoner. We use the classical entailment operator, $\approx\!\!\!\!|$, for statements contained in $\mathcal{K}$. For example, if $\mathcal{K}$ contains the statement α, then it is represented as $\mathcal{K} \mid\!\approx \alpha$. Furthermore, we use a logical entailment operator, $\vdash$, to denote the logical consequence of a set of sentences, A.

2.2 Related Work: Defeasible Reasoning

Defeasible reasoning is an important form of non-monotonic reasoning because it introduces mechanisms for strengthening and weakening new information presented to a reasoner. Defeasible reasoning also deals with atypical reasoning situations in which humans inherently reason differently, whether due to their beliefs, context or other circumstances. Defeasible reasoning and its mechanisms allow AI systems to represent human reasoning more accurately than classical reasoning and more accurately than the basic notion of non-monotonic reasoning. We will illustrate the defeasible reasoning approach with an example.

Example. We use the well-known “Tweety is a bird” example and make a slight adaptation to it. In this example, the reasoner knows that:
(i) Birds typically fly
(ii) Birds have wings
(iii) Penguins are birds
(iv) Tweety is a penguin

Can the reasoner infer that Tweety can fly? From the consequent of (iii), the reasoner can conclude that Tweety is a bird. The reasoner may then infer that Tweety can fly since (i) normally follows from the consequent of (iii). Due to biological reasons, however, we know that penguins cannot fly. Thus, penguins are an exception to typical birds. In this example, the reasoner′s knowledge that birds typically fly is weakened with the additional information made known to the reasoner.

2.3 Related Work: Belief Revision

With belief revision, we consider the beliefs of a reasoner as a belief base and consider additional beliefs as part of some belief set. When a reasoner receives conflicting information, the conflict indicates flawed prior beliefs of the reasoner. As a result, inferences from prior beliefs need to be retracted and the new information is then taken into account.

Example. Consider the following scenario:
(i) Linguists typically speak more than three languages
(ii) Sarah is a linguist

In the case of belief revision, statements (i) and (ii) are in the reasoner's belief set known to the reasoner. Suppose, in addition to (i) and (ii), it is made known to the reasoner that Sarah speaks only two languages. This new information is in conflict with the reasoner's existing beliefs. To resolve the conflict, belief revision seeks to revise the knowledge known to the reasoner.

Alchourrón, Gärdenfors and Makinson (AGM) [2] argue that a reasonable revision function, $*$, must adhere to their proposed 8 postulates for belief revision. The notation $\mathcal{K} * \mu$ denotes revising the belief base, $\mathcal{K}$, with new belief μ. The 8 AGM postulates for belief revision are analysed by Katsuno et al. [15], along with other revision methods.

2.4 Related Work: Belief Update

With belief update, there exist several possible states of the world in which the statements contained in the belief base are true. When a reasoner encounters new information, it is because the world itself has undergone a change and the reasoner will select the closest corresponding state of the world to incorporate this change.

Example. Consider the following scenario:
A room contains two objects in it, a book and a magazine. Suppose b means the book is on the floor and m means the magazine is on the floor. Then, ψ states that either the book is on the floor or the magazine is on the floor, but not both. Now, we order a robot to put the book on the floor. The result of this action should be represented by the update of ψ with b. After the robot puts the book on the floor, all we know is b. It remains open whether the magazine is on the floor or not.

Katsuno and Mendelzon [16] propose 8 postulates which they argue should be satisfied for any belief update function and also contrast their proposed postulates with those of belief revision. They use $\psi \diamond \mu$ to denote the result of updating $\mathcal{K}$ with μ.

3 DEFEASIBLE REASONING

Defeasible reasoning is a form of non-monotonic reasoning used to represent human reasoning. It is well known that humans argue differently and make different conclusions. When said conclusions have the potential to be retracted, based on additional or external information made known to the reasoner, the reasoner is said to reason defeasibly. To model defeasible reasoning, a reasoning language that goes beyond classical reasoning needs to be defined. We describe the theory behind the defeasible language in this section, illustrate its application and lastly describe defeasible properties pertinent to our study.

3.1 Defeasible Logic

When a conclusion is deemed defeasible, it is understood that in typical reasoning circumstances the conclusion is valid. However, additional information made known to the reasoner could either support or invalidate the conclusion. Defeasible reasoning thus allows the conclusion drawn by the reasoner to be strengthened or weakened when additional information is made known to the reasoner. The crux of defeasible reasoning lies in the interpretation of "typical" or "normal" circumstances. When humans reason, they are confronted with diverse reasoning situations: some of which are ordinary and some of which are unusual. Defeasible reasoning accounts for the sometimes unusual nature of human reasoning in that it considers statements which may ordinarily be true, but not always. We attach to the term "typically" or "normally", the meaning that the antecedent is a plausible reason to believe the consequent. The language of defeasible reasoning extends classical reasoning with the defeasible entailment operator, $\mid\sim$, which has the meaning of "typically " as described above. Similar to the example described in the introduction, we consider a general conditional statement of the form "If A is the case, then typically B is the case". "If A is the case" is the antecedent of the conditional and "then typically B is the case" is the consequence. By letting the antecedent be α and the consequent be β, under defeasible circumstances, this is represented as $\mathcal{K} \mid\approx \alpha \mid\sim \beta$. Two particular forms of defeasible reasoning have been studied widely in the AI community: prototypical and presumptive reasoning. Each establishes the typical conditions as perceived by a human reasoner, under which an instance of a category is believed to be true. These will be described in the subsequent sections.

3.1.1 Prototypical Reasoning. Computational models of cognition have been developed to represent common sense reasoning, but with no clear success. Typicality-based reasoning [21, 22] is an approach to represent common sense reasoning by considering instances of a category which are considered better instances than others. With this type of reasoning, a prototype for each reasoning scenario is assumed. The prototype has certain typical features, though these features cannot be thought of as necessary or sufficient conditions [21]. As a brief example, we consider a person establishing the conditions under which a substance is believed to be water. In classical reasoning, the undisputed knowledge known to the reasoner is that the chemical formula for water is H_2O. However, to categorise a sample of water, a reasoner makes uses of the knowledge that water is usually colourless, usually tasteless and usually odourless. Ordinarily, a reasoner might not perform a chemical analysis on the sample of water and might use some of its prototypical features in his reasoning process. The notion of typicality introduced here is similar to that of defeasible reasoning.

3.1.2 Presumptive Reasoning. A presumptive argument is said to be an argument which goes beyond its premises [30]. Presumptive reasoning suggests that an argument may have multiple possible consequences. An example of a presumptive argument used in practice is to consider the prosecution of a suspect, q, based on the testimony of the witness, p. The fact that the witness has testified as such does not logically imply the fact that the suspect was at the crime scene. In particular, when the witness testimony is intentionally false, whether due to inaccurate observations or flawed recollection, the suspect may not have been at the crime scene at all. The argument from p to q is presumptive since p does not logically imply q. The case made against the suspect is thus dependent on valid premises made by all witnesses. As a result, invalid premises could lead to the retraction of previous conclusions. When the argument allows for conclusions to be retracted, the argument is also said to be defeasible. Walton [31] explained how to go about evaluating an argument, in particular when Bayesian systems applied to natural arguments, such as the example above, are not successful.

3.2 Properties

The KLM [17] properties of defeasible reasoning are widely acknowledged in the defeasible reasoning and non-monotonic reasoning community. These properties are pertinent to our study and we discuss them below. For each property, we use the notation for propositional logic described in Section 2.1 and the notation for defeasible logic described in Section 3.1. In addition, we show some properties of interest, derived from the KLM [17] properties, also used in our study.

1. Reflexivity (REF)

$\mathcal{K} \approx \alpha \mid\sim \alpha$

Reflexivity states that if a formula is satisfied, it follows that the formula can be a consequence of itself.

2. Left Logical Equivalence (LLE)

$$\frac{\mathcal{K} \approx \alpha \leftrightarrow \beta, \mathcal{K} \approx \alpha \mid\sim \gamma}{\mathcal{K} \approx \beta \mid\sim \gamma}$$

Left Logical Equivalence states that logically equivalent formulas have the same consequences.

3. Right Weakening (RW)

$$\frac{\mathcal{K} \approx \alpha \rightarrow \beta, \gamma \mid\sim \alpha}{\mathcal{K} \approx \gamma \mid\sim \beta}$$

Right Weakening expresses the fact that one should accept as plausible consequences all that is logically implied by what one thinks are plausible consequences.

4. And

$$\frac{\mathcal{K} \approx \alpha \mid\sim \beta, \mathcal{K} \approx \alpha \mid\sim \gamma}{\mathcal{K} \approx \alpha \mid\sim \beta \wedge \gamma}$$

And expresses the fact that the conjunction of two plausible consequences is a plausible consequence.

5. Or

$$\frac{\mathcal{K} \approx \alpha \mid\sim \gamma, \mathcal{K} \approx \beta \mid\sim \gamma}{\mathcal{K} \approx \alpha \vee \beta \mid\sim \gamma}$$

Or says that any formula that is, separately, a plausible consequence of two different formulas, should also be a plausible consequence of their disjunction.

6. Cautious Monotonicity (CM)

$$\frac{\mathcal{K} \approx \alpha \mid\sim \beta, \mathcal{K} \approx \alpha \mid\sim \gamma}{\mathcal{K} \approx \alpha \wedge \beta \mid\sim \gamma}$$

Cautious Monotonicity expresses the fact that learning a new fact, the truth of which could have been plausibly concluded, should not invalidate previous conclusions.

The properties of **REF, LLE, RW, And, Or** and **CM** are considered the 6 KLM [17] properties of defeasible reasoning and, as such, are pertinent to our study. The defeasible reasoning properties of **Cut, RM, Transitivity** and **Contraposition**, although not KLM [17] properties, are also widely recognised in the defeasible reasoning community and therefore have relevance to this study. We illustrate and describe each of these properties as Supplementary Information in Section 9.1.

4 PROBLEM STATEMENT

In this section, we outline our research questions. We then describe the anticipated outcomes and expected impact of this research.

4.1 Research Questions

RQ1. Do the theoretical descriptions of defeasible reasoning compare normatively or descriptively with human reasoning?

This research question seeks to identify the philosophical link between human reasoning and defeasible reasoning. A normative [3, 28] relationship would suggest that humans reason according to believed norms or standards, which are also generally accepted by other reasoners. In contrast, a descriptive [1, 4, 6, 23] relationship would suggest that humans share the speculative aspect of human reasoning with defeasible reasoning, choosing to consider external sources of information as additional grounds on which to make an inference.

RQ2. Which of the 6 KLM [17] properties of defeasible reasoning are typically employed in human reasoning?

The KLM [17] properties have not been extensively tested with human reasoners, yet they are widely accepted as the standard properties for defeasible reasoning. In ordinary reasoning situations, it remains to be empirically tested whether each property is indeed employed in human reasoning or not.

RQ3. What are the limitations of the KLM [17] defeasible reasoning approach when compared to human reasoning?

This question asks whether every human reasoner can reason defeasibly. Aside from the properties, the mechanisms and language of defeasible reasoning may have inherent differences to human reasoning which need to be isolated and mitigated to represent human reasoning accurately.

RQ4. Which of defeasible reasoning, belief revision or belief update shares more features with and is a better representation of human reasoning?

This question is not answered in this paper, but is left for future work. We propose that humans argue defeasibly and provide support in terms of the number of KLM [17] properties that have correspondence with human reasoning in natural reasoning circumstances. Two other independent studies have been undertaken, each focusing on the formalism of belief revision and belief update respectively. A study, comparing the findings of this study with that of the two studies focusing on belief revision and belief update, is left for future work.

4.2 Anticipated Outcomes

As part of this study, we conducted a survey to establish the correspondence between human reasoning and the formal model of defeasible reasoning, based mainly on the commonly accepted KLM [17] properties. The survey was the main output of this study. We also documented the findings of our survey in this paper. We anticipated that designing appropriate questions would not be straightforward. For defeasible reasoning, there are 6 properties which we needed to consider. Each property had to be isolated and tested in the form of an example. The questions pertaining to these properties would also have had to be clear enough for our participants to understand. Another challenge that we anticipated was that we would not obtain enough participants for this survey. We expected that our chosen survey distribution platform, Mechanical Turk, would provide us with access to at least 30 participants for this survey. We also depended on the Mechanical Turk server and network infrastructure to be fully operational during our project time-span. We also relied on the global community of computer science researchers to be interested and willing to participate in our survey.

4.3 Expected Impact

There is currently an emerging research project, in the defeasible reasoning community, to test whether the normative properties of formal systems of defeasible reasoning are appropriate for representing human reasoning. The most significant impact of the work done in our study was thus to contribute towards this emerging research paradigm. To the best of our knowledge, the assumptions underpinning the formalism of defeasible reasoning have not been sufficiently and empirically tested. Another impact of our work was to determine the extent to which the postulated axioms for the domain of defeasible reasoning indeed were true with human reasoning. Additionally, the postulates for belief revision and defeasible reasoning are similar, as can be seen from their formal properties. While those parallels may hold *formally*, however, there is insufficient evidence to argue that they may hold *empirically*. This study thus provides empirical evidence to support the argument that there is a link between defeasible reasoning and human reasoning. This study also provides a first step in determining the link between defeasible reasoning and other forms of non-monotonic reasoning like belief revision and belief update. Finally, the results of this study is another step towards the greater goal of understanding human cognition and contributes to the body of knowledge around non-monotonic and defeasible reasoning.

5 IMPLEMENTATION

In this section, we describe the design of our experiment. We also describe our implementation strategy and expected challenges. We proceed to document our testing and evaluation strategy and lastly account for the theoretical contributions of this study.

5.1 Experiment Design, Strategy and Expected Challenges

We have conducted an experiment which involved participation from human subjects. Our experiment involved three broad phases: (i) the design of suitable survey questions, (ii) the execution of a survey and (iii) the analysis of the survey results. The survey included 18 questions mainly pertaining to the KLM [17] properties of defeasible reasoning. The 18 questions were divided into 2 categories, namely concrete questions and abstract questions. There were 10 concrete questions and 8 abstract questions. Each question contained a reasoning scenario based on a particular property of defeasible reasoning. The conversions from formal properties to human reasoning scenarios were guided by the usual translations between English and the logics in question. The correspondence between abstract and concrete questions, according to the property tested, is shown in Table 3.
Of the 10 concrete questions, 2 questions related to prototypical and presumptive reasoning respectively. The questions about prototypical and presumptive reasoning did not have corresponding abstract questions as they are not governed by any postulates. Instead, the questions about prototypical and presumptive reasoning were included because of relevance to defeasible reasoning as a whole. Questions 2a to 9a were strictly based on the concrete form of the KLM [17] properties, whilst questions 10a to 17a were based on the abstract form of those properties. The participant was then required, for each question, to state whether he agreed or disagreed with the conclusion and provide a reason for his answer. The reason for the answer given by the respondent was crucial to our study. We have used the participants' reasons to identify the reasoning style used and identified whether the reasons provided by our participants were normative or descriptive.

We have chosen a population sample of 30 respondents for our survey. After considering literature pertaining to similarly designed experiments, we note that this is a relatively large sample size. This number has also been chosen because this project had a time span of approximately 5 months.

We have considered 3 platforms on which to conduct our survey. The first was Mechanical Turk [24]. Turk is an online website, hosted by Amazon, which allows registered users, called Workers, to answer surveys virtually for a monetary reward. A second option was to hold a focus group in a controlled venue, such as a lecture theatre in a university. Each participant in the focus group would be compensated. A third option was to publish our survey on Google Forms and distribute it via channels such as the undergraduate computer science mailing list at the University of Cape Town. Respondents would be given the choice to complete the survey and receive compensation, but co-operation was not guaranteed. After considering the three platforms on which to conduct our survey, we have chosen Mechanical Turk. It provided us with access to a wide range of respondents - all registered users of Mechanical Turk. Another reason to prefer Mechanical Turk was that it provided a more representative sample of respondents as opposed to only university students. This platform also gave us control over who we selected as respondents before the survey was distributed: on Mechanical Turk we had the option to choose, for example, the age group and location of our participants, as well as the number of successful previously-completed tasks taken by our participants on Amazon Turk. The criteria for selecting our participants on Mechanical Turk is discussed in Section 6. Google Forms was our second choice of platform because it did not have the logistical complexity that a focus group had. Participants such as university students, however, might not have been interested in our survey and incentives could not be directly implemented using Google Forms.

In terms of ethical, legal and professional issues, we had identified that human participants were going to be involved, as human responses to our questions were required. We had applied to obtain ethical clearance from the Faculty of Science Human Research Ethics Committee at the University of Cape Town and have since received ethical clearance and permission to conduct this study. Upon taking the survey, participants were first presented with a description of the survey, their role as participants and an informed consent form. Contact details of the researchers were also provided. Before the data-handling, all survey responses were anonymised. To this end, we also did not collect names, cellphone numbers or email addresses from our participants. The only personal contact information we collected from each participant was their Amazon Turk Worker ID. Furthermore, there were no apparent professional issues as this was purely a research project. The intellectual property for the final reports reside with the authors and ownership of the questions used in this survey will be provided to the University of Cape Town. These reports have been made available to the survey respondents through the study website and through the researcher contact information as seen in the survey.

5.2 Testing and Evaluation

The questions developed for our survey were tested through a trial set of surveys conducted using a small sample of people as respondents. This was done to test the process involved in using Amazon's Mechanical Turk as well as to give us insight into how our questions might be received and interpreted. The observations made about using Turk in the trial survey is reported in section 6.1. Additionally, we asked a variety of experts and non-experts to evaluate our survey for coherence, clarity and other desirable characteristics of questions, more examples of which can be found in [18].

As an expert in Artificial Intelligence, knowledge representation and reasoning, we have asked our project supervisor, Professor Thomas Meyer, to evaluate our survey. We have also approached an expert in Psychology and Cognitive Science, at the University of Cape Town, but unfortunately, he was not available to evaluate our survey. The remaining experts who evaluated our survey questions comprised one Master's student in computer science, as well as two computer science Honours students also conducting a study on belief revision and belief update respectively. From a general perspective, this survey was evaluated by an international doctoral student in language and African studies as well as a computer science Honours student not familiar with defeasible reasoning. Participants in the final set of surveys were asked to give overall feedback on the experience. This was included to question, and facilitate their reflection on, the authenticity of their responses, given that people think and act differently in reality. Suggestions of things on which to comment were provided - for example, how concrete they considered our examples to be.

Exploratory data analysis such as correspondence analysis was performed on the final data set. In the case of correspondence analysis, we evaluated if the associations between questions and their answers match overall with what we were expecting based on the formalised reasoning patterns. We compared participant agreement across the survey as a whole, but also across concrete and abstract questions. We compared the correspondence of the reasoning style participants used in concrete questions and in abstract questions too. We sought to identify a normative reasoning trend across all questions and discuss our findings in section 7. Responses to survey questions comprised two parts: a closed aspect (a binary response) and a leading aspect (a reason). For the questions that have a binary response, either 'Yes' or 'No', quantitative measures, such as the mode and hit rate, were used for evaluating its success. In this study, we propose to qualify a hit as an agreement of the participant response with the response expected from the formalism in question. Qualitative analysis, done through the process of coding as outlined by Cresswell [7], was applied to explanations given by participants. Ultimately, the results and evaluations of our study were interpreted to decide whether there is evidence to support the formalism of defeasible reasoning having a normative or a descriptive relationship with human reasoning.

6 SURVEY

Mechanical Turk provides the option for task creators, called Requesters, to use external survey sites, such as SurveyMonkey and Google Forms, and link it to Turk. Initially, we wanted to use SurveyMonkey, but it had several limitations including only 10 questions per survey on their free service. In contrast, the use of Google Forms is free of charge, allows for an unlimited number of questions and has a familiar survey interface. Thus, we chose Google Forms as a platform to collect our survey responses. At the end of each survey, each Worker was provided with a six-digit security code. This code was then submitted by each Worker upon the completion of the HIT to Mechanical Turk, as proof of identity. The subsequent sections will explain the specific design, variables and cost involved for our trial and final surveys.

6.1 Trial Survey

For our trial survey, we submitted our survey as a Human Intelligence Task (HIT) on Mechanical Turk [5, 24]. We required responses from 5 Mechanical Turk Workers for our trial survey. The HIT was published in 1 batch with 5 assignments. During our trial batch, we rejected one response and thus a new assignment had to be issued. This response was rejected because the Worker answered in poor English and failed to provide understandable responses to the survey questions. Each Worker was compensated with R30. As the amount paid to each Worker is accepted by Turk only in dollars, the equivalent amount paid to Turk was \$2.1. The total cost of the trial survey for 5 participants was thus R180, or \$12.6 at the time of payment. To approve the HITs and release payment, we selected the option to delay our auto-approval of HITs to 3 days. This gave us sufficient time to qualify each HIT in terms of completeness, good quality answers and to disqualify duplicate or nonsensical answers. We also specified a survey expiry period of 3 days. We specified that HIT Approval Rate (%) for all Requester's HITs was greater than or equal to 90 for each worker. We specified that the number of approved HITs for each worker should be greater than 0. We also did not restrict Workers by location and allowed all categories of Workers to accept our HIT. We anticipated that the trial survey would take 30 minutes to complete, and selected a generous completion time of 50 minutes on Turk. The average response time for Workers in the trial survey was reported by Turk as 36 minutes and 37 seconds.

6.2 Final Survey

For our final survey, we required responses from 30 participants and published the HIT in 5 batches of 6 participants each. This decision was in part because a batch size of greater than 9 participants carries an extra fee owed to Mechanical Turk. The details for each batch is discussed in the subsequent sections.

6.2.1 Batch 1 of 5. During our first batch, we rejected one response and thus a new assignment had to be issued. This response was rejected because the Worker answered by quoting the information in each question without providing insight into his reasoning. The compensation for each worker was increased from R30, as was the case in the trial, to R40. This was to encourage Workers to pay attention to the survey and provide high-quality answers. As the amount paid to Mechanical Turk was in dollars, the equivalent amount paid to each Worker was \$2.68. The total cost of the batch for 6 participants was thus R298.43 or \$20.10 at the time of payment. We maintained the auto-approval delay of 3 days, as was the case in the trial. We also maintained a survey expiry period of three days. We specified that HIT Approval Rate (%) for all Requester's HITs was greater than or equal to 97 for each worker. This increased from 90 in the trial and was decided upon to improve the overall quality of responses. We specified that the number of approved HITs for each worker should be greater than 0. We chose again not to restrict Workers by location and, in contrast to the trial, only allowed Master Workers to accept our HIT. Master Workers are a subset of the ordinary Worker population on Mechanical Turk and they are vetted, internally by Turk, as reliable and superior Workers. However, the criteria to qualify a Worker as a Master Worker is not made available to the public. The fees owed to Turk were higher than in the trial as the Master Worker criteria carried an additional fee of \$0.134 per Worker. We anticipated that the final survey would take 40 minutes to complete, based on the average response time of 36 minutes and 37 seconds in the trial survey, and selected a generous completion time of 50 minutes on Turk. The average response time for Workers in this batch was reported by Turk as 22 minutes and 40 seconds.

6.2.2 Batch 2 of 5. During our second batch, no responses were rejected. The compensation for each worker was kept at R40. As the amount paid to Mechanical Turk was in dollars, the equivalent amount paid to each Worker was \$2.68. The total cost of the batch for 6 participants was R299.30 or \$20.10 at the time of payment. We maintained the auto-approval delay of 3 days. We also maintained a survey expiry period of three days. The criteria for selecting Workers and the survey completion time was the same as in Batch 1. The average response time for Workers in this batch was reported by Turk as 21 minutes and 28 seconds.

6.2.3 Batch 3 of 5. During our third batch, three responses were rejected. The first response was rejected because the answers provided were vague. The second response was rejected because the participant restated his binary response of 'Yes' or 'No' when prompted for a reason. The third response was rejected because the reasons given did not apply to the questions and some reasons were given in poor English. The compensation and payment details were the same as Batch 2. The auto-approval delay, survey expiry period, criteria for selecting Workers and survey completion time was the same as in Batch 1. The average response time for Workers in this batch was reported by Turk as 29 minutes and 4 seconds.

6.2.4 Batch 4 of 5. During our fourth batch, only 5 of the 6 responses were accepted. The response that was rejected, was a duplicate entry and was only detected at a later stage. As a result, a new HIT was not issued. The duplicate entry was removed and the original entry was kept. The compensation for each worker was kept at R40. As the amount paid to Mechanical Turk was in dollars, the equivalent amount paid to each Worker was $2.68. The total cost of the batch for 6 participants was R301.58 or $20.10 at the time of payment. The auto-approval delay, survey expiry period, criteria for selecting Workers and survey completion time was the same as in Batch 1. The average response time for Workers in this batch was reported by Turk as 33 minutes and 42 seconds. One of the responses in this batch provided some feedback regarding the survey completion time. The participant wrote in the general feedback section that he felt rushed and that the timer should be increased. This was taken into consideration for the final batch.

6.2.5 Batch 5 of 5. During our fifth and final batch, no responses were rejected. The compensation for each worker was kept at R40. The compensation and payment details were the same as Batch 4. The auto-approval delay, survey expiry period and criteria for selecting Workers was the same as in Batch 1. In contrast to the previous batches, the survey completion time was increased from 50 minutes to 60 minutes. This was done to accommodate Workers who require more time to complete the survey. The average response time for Workers in this batch was reported by Turk as 33 minutes and 59 seconds.
In summary, the final survey was conducted over approximately 2 days at a total cost of R1500.19. In total, 29 of the 30 HITs issued were both verified and accepted.

7 RESULTS AND DISCUSSION

Our results and discussion are organised as follows. First, we explain our assumptions and give an overview of the data we collected. We then explain how we processed and categorised our responses. Following that, we explain our descriptive statistics for this study. Lastly, in our analysis, we provide key insights from our study.

7.1 Assumptions and Data Overview

In our collected data, the 'Yes' and 'No' responses were considered quantitative data whilst the reasons given by participants were considered qualitative data. We assumed that the property of **Ref** holds for all human reasoners and therefore it was not tested in our study. Upon receiving feedback from our supervisor, it was found that question 6a and 7a were not appropriate models of the properties they intended to test - **RW** and **And** respectively - because the conclusions in those questions contained the condition 'some' which could not be converted to a propositional setting and thus could not be converted to their respective KLM [17] properties. The responses to questions 6a and 7a were removed from our final dataset and removed from our analysis further in this study. As a result, our dataset contained responses from 29 participants to 16 questions only. This gave a total of 464 closed responses and 464 leading responses.

7.2 Response Categories

After considering the reasons provided for the leading responses from our participants, we have identified four main themes: 'Support', 'Speculative', 'Technical' and 'Other'.
'Support' refers to a reason which is fully based on the given information. In this case, the participant chose to believe the given information and quoted it as his response.
'Speculative' refers to a reason in which there is partial support from the given information, but also in which external information, not present in the question, is considered in the reason. In this case, a participant would cite his inherent knowledge or other sources of knowledge is his reason.
'Technical' refers to a reason which has support from the given information in the question, but differs in the meaning of the keyword in the question. For example, a response may contain the phrase "typically, but not always". Keywords and phrases such as "typically" and "A implies B" were identified and explained in the background section of the survey.
'Other' refers to reasons which did not fit into any of the above categories. In this case, the reasons given were not explanatory, not clear or undecided. After identifying these four themes, we have qualified a normative response as being either 'Support' or 'Technical'. We have qualified a descriptive response as only being Speculative. The remaining responses ('Other') were not classified further.

7.3 Descriptive Statistics

In Table 1, we show the descriptive statistics of our processed quantitative data. In Table 2, we show the descriptive statistics of our processed qualitative data.

Table 1: Descriptive statistics for quantitative data

Statistic	**Yes**	**No**
Mean	15,88	13,13
Median	16,5	12,5
Mode	21	8
Standard Deviation	7,11	7,11
Mean Standard Error	1,78	1,78

Table 2: Descriptive statistics for qualitative data

Statistic	**Normative**	**Descriptive**	**Other**
Mean	23,13	5,88	1
Median	23	5	0,5
Mode	23	4	0
Standard Deviation	4,65	3,93	1,51
Mean Standard Error	1,16	0,98	0,38

From these statistics in Table 2, we observe that the mean number of 'Yes' responses (15,88) recorded was 17.32% higher than the mean number of 'No' responses recorded (13,13), and accounted for 54,74% of the total closed responses. From the statistics in Table 3, wee also found that the mean number

Figure 1: Quantitative responses across all abstract and concrete questions

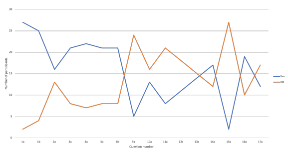


Figure 2: Quantitative responses for defeasible reasoning properties tested in concrete questions

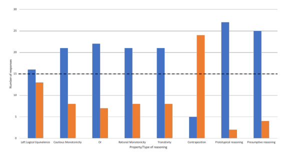


Figure 3: Quantitative responses for defeasible reasoning properties tested in abstract questions

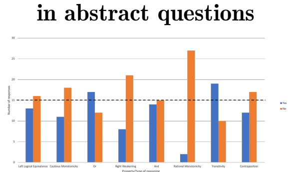


Figure 4: Themes identified in qualitative responses for each defeasible reasoning property across concrete questions

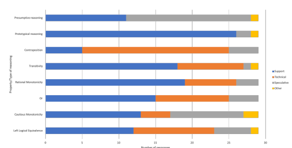


Figure 5: Themes identified in qualitative responses for each defeasible reasoning property across abstract questions

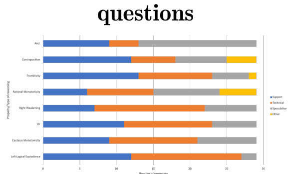
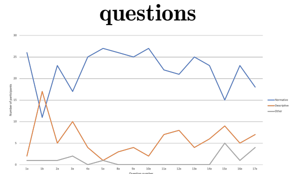


Figure 6: Qualitative responses across all concrete and abstract questions



of 'Normative' responses was 74,58% higher than the mean of the 'Descriptive' responses, which accounted for 76,29% of the total leading responses.

7.4 Analysis

One of our research questions involves determining which of the KLM [17] properties were true with human reasoning. The baseline for our analysis was is calculated at the ceiling of 50% of the total number of responses obtained. The results obtained were considered significant if responses exceeded the baseline amount. A positive hit rate for a particular property indicates that the number of 'Yes' responses exceeded the baseline whilst the number of 'No' responses were below the baseline. In contrast, a negative hit rate for a particular property indicates that the number of 'No' responses exceeded the baseline whilst the number of 'Yes' responses were below the baseline.

Figure 7: Qualitative responses for each defeasible reasoning property across concrete questions

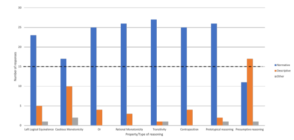
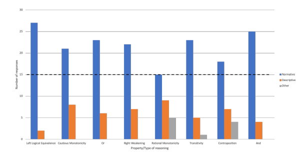


Figure 8: Qualitative responses for each defeasible reasoning property across abstract questions



7.4.1 Insight 1. In Figure 2, we observe that the hit rate for the concrete forms of prototypical and presumptive reasoning respectively is positive. A very strong majority of participants have agreed with the proposed conclusion. As can be seen in 1, for question 1a, 27 participants said 'Yes' whilst only 2 said 'No' and for question 1b, 25 participants said 'Yes' whilst only 4 said 'No'. From the positive hit rate, we found that humans can reason both prototypically and presumptively.

7.4.2 Insight 2. In Figure 2 and 3, we observe that the hit rate for the concrete and abstract forms of the property **Or**, is positive. A strong to moderate majority of participants agreed with the proposed conclusion. As can be seen in 1, for question 4a, 22 participants said 'Yes' whilst only 7 said 'No' and for question 14a, 17 participants said 'Yes' whilst only 12 said 'No'. From the positive hit rate, we found that humans can reason according to the Kraus-Lehmann-Magidor [17] property of **Or**.

7.4.3 Insight 3. In Figure 2 and 3 respectively, we observe that the hit rate for the concrete and abstract forms of the property **Transitivity**, is positive. A strong majority of participants agreed with the proposed conclusion. As can be seen in 1, for question 5a, 21 participants said 'Yes' whilst only 8 said 'No' and for question 14a, 19 participants said 'Yes' whilst only 10 said 'No'. From the positive hit rate, we found that humans can reason according to the property of **Transitivity**.

7.4.4 Insight 4. In Figure 2 and 3 respectively, we observe that the hit rate for the concrete and abstract forms of the property **Contraposition**, is negative. A very strong to moderate majority of participants disagreed with the proposed conclusion. As can be seen in 1, for question 9a, only 5 participants said 'Yes' whilst 24 said 'No' and for question 17a, only 12 participants said 'Yes' whilst 19 said 'No'. From the negative hit rate, we found that humans do not reason according to the property of **Contraposition**.

7.4.5 Insight 5. For the concrete and abstract properties of **LLE, CM** and **RM**, opposing majorities of responses were obtained. In Figure 2, the hit rate for the concrete form of **LLE** is moderately positive whereas the abstract form of **LLE** has a moderately negative hit rate, which can be seen in

Figure 3. In Figure 2, the hit rate for the concrete form of **CM** is strongly positive whereas the abstract form of **CM** has a strongly negative hit rate, which can be seen in Figure 3. In Figure 2, the hit rate for the concrete form of **RM** is strongly positive whereas the abstract form of **RM** has a strongly negative hit rate, which can be seen in Figure 3. From the discrepancy in hit rates across abstract and concrete questions for the properties of **LLE, CM** and **RM**, it is inconclusive as to whether there is a significantly positive or negative relationship between human reasoning and these defeasible reasoning properties.

7.4.6 Insight 6. n Figure 4, we show the prevalence of each of the four themes for the KLM [17] properties across the concrete questions. In Figure 5, we show the prevalence of each of the four themes for the KLM [17] properties across the abstract questions. An increased number of 'Technical' responses from concrete questions (Figure 4) to abstract questions (Figure 5) can be seen. We have identified that participants strongly motivated their responses using only the information provided in the question for the concrete questions, whilst for abstract questions, participants motivated with an increased number of 'Technical' and 'Speculative' responses.

7.4.7 Insight 7. In Figure 6, we observe that across all abstract and concrete questions, the majority of participants tended to reason normatively. In a small number of abstract questions, notably in 15a and 17a, a growing number of participants were not able to reason normatively or descriptively. In these questions, the reasons became vague and we observed that participants became tired and confused towards the end of this survey, potentially resulting in an increased number of speculative or poorly thought out answers. In question 1b, we observe the only instance in which more participants reasoned descriptively than normatively. The question involved deals with presumptive reasoning and asks, given certain information, whether magic penguins typically fly. This question deals with a fictitious scenario which might have confused participants and thus resulted in more speculative answers as opposed to normative answers.

7.4.8 Insight 8. After analysing the responses further, we have found a positive hit rate for normative reasoning in both the concrete and abstract forms of the defeasible reasoning properties. We show the hit rate for reasoning style used when properties were presented in concrete form, in Figure 7, and abstract form, in Figure 8. In the case of abstract defeasible reasoning properties, the hit rate for **RM** was equivalent to the baseline. However, the majority of participants, as compared to the other reasoning styles for **RM**, still reasoned normatively.

From Insights 7 and 8, we found that humans have a normative relationship with defeasible reasoning when properties are presented in both concrete and abstract form. However, more humans reason descriptively when faced with abstract reasoning situations than when faced with concrete reasoning situations.

8 CONCLUSIONS

It is well-known that humans reason non-monotonically. Various formalisms of non-monotonic reasoning have emerged in the AI community and this study focused on identifying and reporting on the relationship between human reasoning and the theoretical descriptions of one of the forms of non-monotonic reasoning called defeasible reasoning. We conducted a survey, containing questions about concrete and abstract reasoning scenarios, with human participants. This study aimed to obtain empirical results to support our hypothesis that there exists a relationship between human reasoning and defeasible reasoning. Each of the questions was based on the KLM [17] and related properties of defeasible reasoning. Each property was formulated as two questions: one with the property in abstract form and the other in concrete form. We now mention some of the important findings from our study. From our data and analyses, we found that humans were able to reason both prototypically and presumptively. We were also able to conclude that humans were able to reason according to the KLM [17] property of **Or** and according to the property of **Transitivity**. We were able to conclude that humans did not reason according to the property of **Contraposition**. For the concrete and abstract KLM [17] properties of **LLE, CM** and **RM**, discrepancies between hit rates existed. It was thus inconclusive as to whether there was a positive or negative relationship between human reasoning and these defeasible reasoning properties. For the concrete and abstract properties of **And** and **RW**, responses could not be used due to the questions not adhering to the propositional nature of the properties, through the use of the keyword 'some'. Further research needs to be done to establish the relationship between human reasoning and the defeasible properties of **And** and **RW** in concrete form. We were able to answer our research question identifying the philosophical link between human reasoning and defeasible reasoning. We found that humans have a normative relationship with defeasible reasoning when properties are presented both in concrete and abstract form. However, more humans reason descriptively when faced with abstract reasoning situations than when faced with concrete reasoning situations.

Ideally, every human reasoner would have the same understanding of the notion of typicality and would have the same knowledge base. Our findings show that this was not the case. In our study, there were discrepancies between survey responses choosing to believe only the given information and choosing to introduce and consider new external information. Further work should be done to improve the understanding of how humans reason in typical reasoning scenarios. To this end, the sample population should extended to include human reasoners, across all ages and across all knowledge levels, to identify discrete differences in human and defeasible reasoning.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Professor Thomas Meyer, for his assistance and feedback on this project. I would also like to thank my two colleagues, Claire and Paul.

REFERENCES

[1] A. Aamodt and E. Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications* 7, 1 (1994), 39–59. https://doi.org/10.3233/AIC-1994-7104

[2] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic* 50, 2 (1985), 510–530. https://doi.org/10.2307/2274239

[3] A. Artosi, P. Cattabriga, and G. Governatori. 1994. An automated approach to normative reasoning. (1994), 132–145.

[4] T.R. Besold and S.L. Uckelman. 2018. Normative and descriptive rationality: from nature to artifice and back. *Journal of Experimental Theoretical Artificial Intelligence* 30, 2 (2018), 331–344. https://doi.org/10.1080/0952813X.2018.1430860

[5] M. Burhmester, T. Kwang, and S.D. Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5. https://doi.org/10.1177/1745691610393980

[6] J.S. Carroll, M.H. Bazerman, and R. Maury. 1988. Negotiator cognitions: A descriptive approach to negotiators' understanding of their opponents. *Organizational Behavior and Human Decision Processes* 41, 3 (1988), 352–370.

[7] J.W. Cresswell. 2014. *Research design: qualitative, quantitative, and mixed methods approaches* (4 ed.). SAGE Publications, Thousand Oaks, California.

[8] M. Denecker. 2000. Extending classical logic with inductive definitions. *In International Conference on Computational Logic* (July 2000), 703–717. https://doi.org/10.1007/3-540-44957-4_47

[9] M. A. Falappa, G. Kern-Isberner, and G. R. Simari. 2002. Explanations, belief revision and defeasible reasoning. *Artificial Intelligence* 141 (2002), 1–28. https://doi.org/10.1016/S0004-3702(02)00258-8

[10] D.M. Gabbay. 1982. Intuitionistic basis for non-monotonic logic. *In International Conference on Automated Deduction* (June 1982), 260–273. https://doi.org/10.1007/BFb0000064

[11] D.M. Gabbay. 1985. Theoretical foundations for non-monotonic reasoning in expert systems. *In Logics and models of concurrent systems* (1985), 439–457. https://doi.org/10.1007/978-3-642-82453-1_15

[12] P. Gärdenfors and D. Makinson. 1994. Nonmonotonic inference based on expectations. *Artificial Intelligence* 65, 2 (1994), 197–245.

[13] G. Gottlob, F. Scarcello, and M. Sideri. 2002. Fixed-parameter complexity in AI and nonmonotonic reasoning. *Artificial Intelligence* 128, 1-2 (2002), 55–56. https://doi.org/10.1016/S0004-3702(02)00182-0

[14] H. Katsuno and A. O. Mendelzon. 1989. A Unified View of Propositional Knowledge Base Updates. *Artificial Intelligence* 11, 2 (1989), 1413–1419.

[15] H. Katsuno and A. O. Mendelzon. 1991. Propositional knowledge base revision and minimal change. *Artificial Intelligence* 3 (12 1991), 263–294.

[16] H. Katsuno and A. O. Mendelzon. 2003. On the Difference between Updating a Knowledge Base and Revising it. *Belief Revision* 29 (2003), 183.

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

[18] J. Krosnich and S. Presser. 2009. Question and Questionnaire Design. *Handbook of Survey Research* (03 2009).

[19] D. Lehmann. 1998. Stereotypical reasoning: Logical Properties. *Logic journal of the IGPL* 6, 1 (1998), 49–58. https://doi.org/10.1093/jigpal/6.1.49

[20] D. Lehmann and M. Magidor. 1992. What does a conditional knowledge base entail? *Artificial Intelligence* 55, 1 (1992), 1–60. https://doi.org/10.1016/0004-3702(92)90041-U

[21] A. Lieto, A. Minieri, A. Piana, and D.P. Radicioni. 2015. A knowledge-based system for prototypical reasoning. *Connection Science* 27, 2 (2015), 137–152. https://doi.org/10.1080/09540091.2014.956292

[22] A. Lieto, D. Radicioni, V. Rho, and E. Mensa. 2017. Towards a unifying framework for conceptual represention and reasoning in cognitive systems. *Intelligenza Artificiale* 11, 2 (2017), 139–153. https://doi.org/10.3233/IA-170111

[23] E.H. Mandani. 1976. Application of fuzzy logic to approximate reasoning using linguistic synthesis. In *Proceedings of the sixth international symposium on Multiple-valued logic*. IEEE Computer Society Press, 196–202.

[24] G. Paolacci and J. Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science* 25 (2014), 184–188.

[25] F. Pelletier and R. Elio. 2005. The Case for Psychologism in Default and Inheritance Reasoning. *Synthese* 146 (08 2005), 7–35. https://doi.org/10.1007/s11229-005-9063-z

[26] J. Pollock. 1991. A Theory of Defeasible Reasoning. *International Journal of Intelligent Systems* 6 (1991), 33–54.

[27] M. Ragni, C. Eichhorn, T. Bock, G. Kern-Isberner, and A. Tse. 2017. Formal Nonmonotonic Theories and Properties of Human Defeasible Reasoning. *Minds and Machines* 27 (2017), 79–117. https://doi.org/10.1007/s11023-016-9414-1

[28] L.W. Van Der Torre and Y.H. Tan. 1999. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law* 7, 1 (1999), 51–67. https://doi.org/10.1023/A:1008359312576

[29] J. van Benthem. 2007. Dynamic logic for belief revision. *Journal of Applied Non-classical Logics* 17, 2 (2007), 129–155. https://doi.org/10.3166/jancl.17.129-15

[30] B. Verheij. 2016. Correct grounded reasoning with presumptive arguments. *European Conference on Logics in Artificial Intelligence* (2016), 481–496. https://doi.org/10.1007/978-3-319-48758-8_31

[31] D. Walton. 2016. Some artificial intelligence tools for Argument evaluation: An introduction. *Argumentation* 30, 3 (2016), 317–340. https://doi.org/10.1007/s10503-015-9387-x

[32] R.R. Yager. 1997. Fuzzy logics and artificial intelligence. *Fuzzy sets and systems* 90, 2 (1997), 193–198. https://doi.org/10.1016/S0165-0114(97)00086-9

9 SUPPLEMENTARY INFORMATION

9.1 Other Defeasible Reasoning Properties

The defeasible reasoning properties, other than the 6 KLM [17] properties, that are relevant to our study are:

A. Cut

$$\frac{\mathcal{K} \approx \alpha \wedge \beta \mid\!\sim \gamma,\ \mathcal{K} \approx \alpha \mid\!\sim \beta}{\mathcal{K} \approx \alpha \mid\!\sim \gamma}$$

It expresses the fact that one may, in his way towards a plausible conclusion, first add an hypothesis to the facts he knows to be true and prove the plausibility of his conclusion from this enlarged set of facts and then deduce (plausibly) this added hypothesis from the facts.

B. Rational Monotonicity (RM)

$$\frac{\mathcal{K} \approx \alpha \mid\!\sim \gamma,\ \mathcal{K} \not\approx \alpha \mid\!\sim \neg\beta}{\mathcal{K} \approx \alpha \wedge \beta \mid\!\sim \gamma}$$

Rational Monotonicity expresses the fact that only additional information, the negation of which was expected, should force us to withdraw plausible conclusions previously drawn.

C. Transitivity

$$\frac{\alpha \mid\!\sim \beta,\ \beta \mid\!\sim \gamma}{\alpha \mid\!\sim \gamma}$$

Transitivity expresses that if the second fact is a plausible consequence of the first and the third fact is a plausible consequence of the second, then the third fact is also a plausible consequence of the first fact.

D. Contraposition

$$\frac{\alpha \mid\!\sim \beta}{\neg\beta \mid\!\sim \neg\alpha}$$

Contraposition allows the converse of the original proposition to be inferred, by the negation of terms and changing their order.

9.2 Correspondence Between Questions In Survey

In Table 3, we show a table comparing the questions testing the same properties of defeasible reasoning. We also distinguish between the different forms of properties, viz. concrete or abstract.

Table 3: Related questions and properties

Property/Type	Concrete	Abstract
Prototypical reasoning	1a	-
Presumptive reasoning	1b	-
LLE	2a	10a
CM	3a	12a
Or	4a	14a
Transitivity	5a	16a
RW	-	11a
And	-	13a
RM	8a	15a
Contraposition	9a	17a

9.3 Raw Data

In Tables 4 and 5, we provide the responses as given by our participants in the survey. We separate quantitative and qualitative data and provide totals for each column.

Table 4: Quantitative data

Question	Yes	No
1a	27	2
1b	25	4
2a	16	13
3a	21	8
4a	22	7
5a	21	8
6a*	9	20
7a*	28	1
8a	21	8
9a	5	24
10a	13	16
11a	8	21
12a	11	18
13a	14	15
14a	17	12
15a	2	27
16a	19	10
17a	12	17
Totals	**254**	**210**

Table 5: Qualitative data

Question	Normative	Descriptive	Other
1a	26	2	1
1b	11	17	1
2a	23	5	1
3a	17	10	2
4a	25	4	0
5a	27	1	1
6a*	25	4	0
7a*	24	5	0
8a	26	3	0
9a	25	4	0
10a	27	2	0
11a	22	7	0
12a	21	9	0
13a	25	4	0
14a	23	6	0
15a	15	9	5
16a	23	5	1
17a	18	7	4
Totals	**354**	**94**	**16**

9.4 Participant Feedback

In Figures 9, 10, 11, 12 and 13, we provide the responses to the feedback questions, as given by the participants in our survey.

Figure 9: Participant age

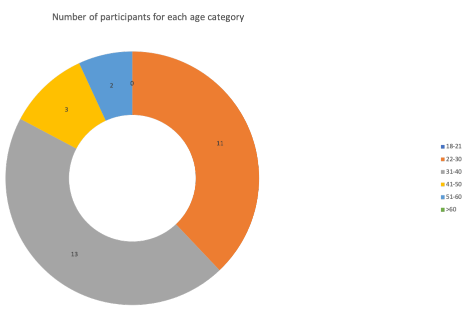


Figure 10: Participant interest in survey

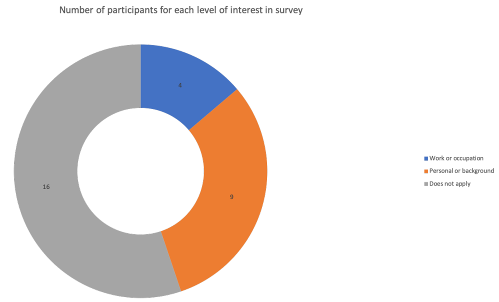


Figure 11: Clarity of survey instructions

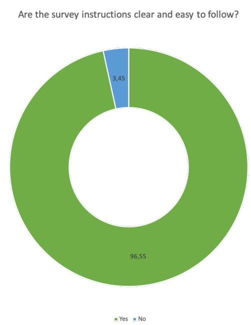


Figure 12: Natural quality of concrete questions

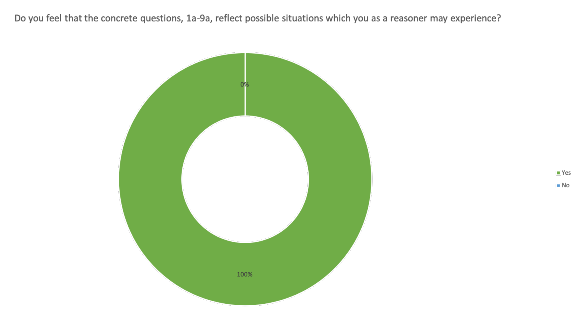


Figure 13: Clarity of abstract questions