Developing a Tool for Eliciting Users' Moral Theories for Automated Moral Agents

By

Kyle Seakgwa
SKGKYL001

Supervisor: Maria Keet

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN

In partial fulfilment of the requirements for the degree

MASTERS IN INFORMATION TECHNOLOGY

Department of Computer Science
UNIVERSITY OF CAPE TOWN

COPYRIGHT INFORMATION

DECLARATION

I, Kyle Seakgwa, hereby declare that the work on which this dissertation/thesis is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature:
Date: 12 February 2024 .....

ABSTRACT

In recent work, Rautenbach and Keet have developed a model of a system, which they name Genet, that allows the user to choose which moral theory their automated moral agent will follow. What remains unclear, however, is how the users will make this choice, given that most of them will not have the vocabulary to classify themselves in the moral philosophical terms used by Genet. This issue is what this thesis is meant to address. This was done by building three high fidelity prototypes and then conducting online user evaluations of them. Each of these prototypes implemented an algorithm that was designed based on the elicitation approach of one of three fields: cognitive science, human computer interaction and knowledge engineering. Each of these aimed to computationally determine a user's preferred moral theory, by availing itself of a human-in-the-loop component enabled by discipline specific elicitation stimuli and rules to classify the user. These prototypes were then evaluated from a usability perspective using the System Usability Scale (SUS), and from an accuracy perspective, to determine which is most validly able to elicit users' moral preferences in the form required by Genet. This latter evaluation was done using validation measures based on existing approaches to validation in moral psychology.

It was observed that all the prototypes performed equally well in terms of usability, with each having an acceptable SUS score. However, each of the prototypes also performed equally inaccurately in terms of the validity of the moral theory categorizations made. While this evaluation was carried out with only a small sample size (n=20) and thus has limited generalizability, as the first study to compare and computationally implement different moral theory elicitation approaches, the present study contributes to evidence for (or at least fails to falsify) problems with the project of making the design of Automated Moral Agents dependent on elicitation of a user's one preferred moral theory. A positive claim that the data collected here does support is that, at least for some potential users, even computational elicitation tools that use empirically validated measures of moral theory preferences (like those from cognitive science) do not allow one to predict the moral judgements they will make.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Associate Professor Maria Keet, for her diligent guidance this long journey. I would also like to thank my wife, Robyn Pasensie, for her unwavering support and patience during the completion of this project. Finally, I would like to thank my mother, Elaine Wolfe, for helping me get to this point.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AMA: Automated Moral Agent
CA: Cronbach's Alpha
DCT: Divine Command Theory
ES: Egoism Scale
MFDA: Moral Foundations base on Divine Authority
MFQ: Moral Foundations Questionaire
MPS/Q: Moral Philosophical Statement/Question
OUS: Oxford Utilitarianism Scale
OUS-IH: Oxford Utilitarianism Scale – Instrumental Harm
OUS-IB: Oxford Utilitarianism Scale – Impartial Beneficence
TE: Thought Experiment
VSD: Value Sensitive Design
VS: Value Scenario

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The ability to distinguish and decide between morally permissible and impermissible actions is something that humans regularly employ [1]. Recent research (e.g., [2][3][4]) has started attempting to impart various computational devices with this capacity as well, thus creating automated moral agents (AMAs). Most of this research has concentrated on designing AMAs that make moral decisions based on a single moral theory e.g., [2], which can be thought of as a framework for deciding what makes something right or wrong. Sometimes no moral theory guides the AMA and the system is instead designed to maximize some amoral value/goal (e.g., stakeholder happiness)[5]. However, single theory approaches can lead to issues, such as when an action taken by the AMA (according to its moral theory or amoral goal) diverges from the desires of the user(s).

Rautenbach and Keet [1] have designed a model to provide an AMA with the capacity for the user to choose its moral theory. They named this model Genet [1]. It models moral theories as having three-levels. The first level is the most general class of moral theory, the second level is a child class which inherits from the moral theory class called the "base theory" [1, p. 3], and the third level is the instantiation of this base theory, known as the "theory instance" [1, p.3]. There are user-configurable attributes at both the base theory and theory instance level [1].

Although this model allows users to configure its moral theory on both levels, it is unclear how to allow the user to actually do this [1]. This is because most humans are unfamiliar with the language and concepts usually used in academic moral philosophy to describe various moral theories, which were also used in the design of the Genet model [1]. So, it is unlikely that the user would be able to classify themselves in the terms used by Genet. This process of learning the user's preferred moral theory even with them not having the conceptual or linguistic vocabulary to describe it is sometimes called "theory extraction" [1, p.10], although this thesis will use this term and "theory elicitation" interchangeably. The theory elicitation issue is one that any AMA model that will allow user selected moral theories, not just Genet, will face. It is this issue that this thesis will address by contributing to the development of an elicitation tool that could be used by Genet-based system, or other AMAs with user-configurable moral theories.

1.1 Research questions

The questions this research will tackle are:

1) Which of the three approaches to user preference elicitation (that of HCI, cognitive science, knowledge engineering/elicitation) as instantiated in high fidelity prototypes delivers the highest validity and reliability in its ability to elicit users' moral beliefs at both the level of base-theory and theory instance?
2) Which of the three approaches to user preference elicitation (that of HCI, cognitive science, knowledge engineering/elicitation) as instantiated in high fidelity prototypes is the most highly rated in terms of usability?

1.2 Hypotheses

H1: There is a significant difference between the computational instantiations of psychometric scale-based approaches which are explicitly formulated in analytic moral philosophical terms (representing cognitive science), psychometric scale-based approaches which take values as their unit of analysis when delivered via value scenarios (representing HCI), and proximity scaling approaches delivered via thought experiments (representing knowledge engineering) with regard to how accurately they are able to categorize users according to their moral theory.

H2: There is a significant difference between the computational instantiations of, psychometric scale-based approaches which are explicitly formulated in analytic moral philosophical terms (representing cognitive science), psychometric scale-based approaches which take values as their unit of analysis when delivered via value scenarios (representing HCI) and proximity scaling approaches delivered via thought experiment (representing knowledge engineering) with regard to usability.

2. LITERATURE REVIEW

This chapter reviews research relevant to the development of a moral theory elicitation tool for a system instantiating the Genet model. Here the attention will be on potentially relevant work in four subfields: machine ethics, the cognitive sciences of morality, knowledge elicitation and HCI. These fields were selected as they each deal with either elicitation of information from end users generally, or the elicitation of moral beliefs or values more specifically.

Note that this project, and thus this literature review, will only deal with the question of how to ascertain the base theory that a user subscribes to. The project will thus not deal with approaches to eliciting information required by instance level settings in the Genet architecture.

The review below starts by stating the desiderata for an approach to the design of a theory extraction tool for Genet.

2.1 Desiderata for an approach to theory extraction for the Genet architecture

Before we begin assessing available approaches for their applicability to designing a Genet belief extraction tool, we must decide what is required of such a tool. Below is a list of properties which an approach must have to be a viable basis for the design of an extraction tool for Genet. Brief justifications for each desideratum are included:

2.1.1 *The tool should validly and reliably classify users according to their preferred moral theory:*

This is necessary because the function of the tool is to allow an AMA to behave according to the user's preferred moral theory. Thus, the tool should accurately measure the user's actual moral theory (i.e. its classifications should be valid [6]). The tool should also return similar results across multiple interactions with the same user (i.e. the classifications should be reliable [6]).

2.1.2 *The tool should provide a robust classification even beyond the stimuli used:*

This is necessary because the tool is supposed to enable an AMA to behave according to the moral theory of the user, even in instances that go beyond material used by the tool to elicit the user's moral judgements for classification. Thus, the accuracy of the classification cannot be bound exclusively to the material the tool uses for classification.

2.1.3 *The tool should be usable without previous training:*

Since the tool must elicit moral information from users whether or not they have training or familiarity with the categories of analytic moral philosophy, the tool should not presuppose or require such training or familiarity. In order to ensure the tool is accessible and usable by as wide a user base as possible, the tool should also require any other special training to properly interact with it.

2.1.4 *The tool should use simple and unambiguous elicitation stimuli and/or terminology:*

In order to ensure the accessibility and useability referred to in section 2.1.1, the tool should use simple stimuli and/or terminology. This is also necessary because the simpler the elicitation stimuli, the easier it is to know which features of the stimuli are driving the users' judgements. This in turn makes it easier to ensure the measurements are valid, and not tracking irrelevant

features of the stimuli.

2.1.5 *The stimuli used should not include features that are not explicitly measured:*

Since, as set out in section 2.1.1, the tool must validly and reliably elicit the user's preferred moral theory, the stimuli used in the tool must not introduce factors that may distort their moral judgement (e.g. if the a tool uses moral scenarios and a scenario includes the killing a person, the presence or hints of information about his/her status should be avoided, the status include young/old, male/female, rich/poor, member of large/small family, race, type of work, health status, presence of criminal charges and so on).

2.1.6 *The tool must capture information that can be mapped onto the categories of analytic moral philosophy:*

This is necessary since the dimensions of an ethical theory (e.g., relevant patients, being consequentialist or non-consequentialist etc) that Genet uses as the essential attributes for representing ethical theories in general are usually associated with analytic ethics and metaethics (e.g., [7]). As a result, approaches which result in outputs that do not share the essential features of analytic moral theories could be difficult or impossible for Genet to model. For instance, some approaches in moral psychology which are most similar to virtue ethics, an ethical view that has no explicit principles, would be difficult for Genet to model [7].

2.1.7 *The tool should be suited to extracting multiple different theories:*

The key feature of Genet is that it allows the user to select their preferred moral theory [1]. Thus, it is essential that the theory extraction tool for Genet can capture multiple theories. Therefore, the approach taken to designing it needs to facilitate this.

2.1.8 *The tool must be automated:*

Genet was conceived as a way to enable automated moral decision making according to a moral theory of the users' choosing, as an alternative to other automated systems only equipped to make moral decisions according to one moral theory [1]. To introduce human assistance (outside of information provided by the user) in the theory extraction stage of the process wouldthus compromise the vision of a system which does what one-theory automated moral reasoning systems do, but with more customizability.

2.1.9 *The tool must be relatively easily extensible:*

To allow the user to select their preferred moral theory in a case where the user's moral theory is not already represented in the system, Genet will have to make adding the theory, as well as the means for extracting such a theory in future, relatively easy [1].

Desiderata 2.1.1 – 2.1.5 cover issues that any approach to moral elicitation from lay populations must satisfy. Thus, since the literature to be reviewed was intentionally selected because it deals with elicitation from lay populations or at least uses techniques that can also be used for lay populations, these desiderata having been satisfied is already ensured as part of the

literature selection. When evaluating the selected literature, this chapter will thus only focus on desiderata 2.1.6 – 2.1.9.

It should also be noted that the following survey of the literature does not take into consideration recent work on reinforcement learning in machine ethics which breaks from the tradition of implementing moral theories such as [8], as well as work on having AMAs learn norms rather than ethical theories [9]. This is because, while these moral theory-agnostic approaches offer a potentially fruitful route to designing AMAs, they still leave open how to deal with issues such as the AMA behaving in a way the user does not agree with. Thus, there is still a need to map moral information drawn from the users into particular moral theories, to allow AMAs to behave accordingly. Since theory-agnostic approaches eschew reliance on moral theory entirely, they offer no clear guidance for meeting desiderata 2.1.6 and 2.1.7.

The rest of this chapter surveys the relevant literature, beginning with the work of machine ethicists.

2.2 Elicitation in Machine ethics

At present, there is little empirical or empirically-informed research in machine/computer ethics which addresses how computational systems should elicit users' ethical beliefs.

However, there are exceptions such as Awad et al.[10] and their "Moral Machine experiment".They deployed a web-based serious game where users were shown a number of images wherea self-driving car is on a collision course with one of two groups of pedestrians. Users must then push a button if they want the car to change course and hit the other pedestrians instead. This approach thus makes use of moral dilemmas. These are situations where one must choosebetween two morally sub-optimal options. This study got millions of participants from all around the world to take part. It was observed that moral judgment differed along cultural/national lines [10].

Since this approach is already fully automated and online, it clearly meets the automation desideratum for the Genet extraction tool. The popularity of the game/study suggests that it is a delivery method that people find appealing, but the project has yet to be evaluated from a usability perspective.

The stimuli used in [10] to elicit participants' moral judgments also do not map user moral judgment onto the categories Genet employs [7]. Such a mapping seems unlikely to be possible since the 9-factor analytical framework used does not correspond with analytic moral philosophical theories [7].

Another issue is that some recent studies, such as [11], cast doubt on the sacrificial dilemma approach when used alone. The implication of the results in [11] is that there may be people who do not respond to these dilemmas as utilitarians are expected to, but, when given other kinds of questions, respond as utilitarians would [11]. So, using such an approach, which [11] does, would lead to misclassifications [11].

The moral machine approach also faces the issue of the difficulty of pictorially representing abstract concepts. For instance, it is unclear how one would pictorially represent something like adherence to divine command theory. Thus, it is not clear how one would use this approach to extract multiple different theories, as the Genet extraction tool is required to. This also makes it difficult to see how one would implement this approach extensibly.

Another machine ethicist who has investigated peoples' moral/philosophical beliefs is Barger [12] who developed the "Philosophical inventory" [12, p. 1]. This is a questionnaire which categorizes users based on their moral/philosophical outlook. This empirically validated scale

classifies participants as either pragmatic, idealist, or existential in philosophical/moral outlook [12].

Developing such a psychometric questionnaire requires one to be clear about what the items are and how they will be scored. Given this clarity, it should be easy to implement this logic computationally.

The questionnaire approach also can extract multiple theories. This is because what it can extract is bounded only by the content of the items chosen to comprise it.

However, [12] also has the problem of the categories used not mapping onto those Genet uses. Here this issue is particularly severe, since the categories used are quite idiosyncratic (e.g., drawn from continental, analytic and pragmatic philosophy) in comparison to those analytic philosophers use, which Genet also uses. There is also often controversy about translating across these traditions [13].

The final consideration regarding using the approach in [12] is its extensibility. Since the psychometric properties, such as validity and reliability, are important to ensuring a psychometric questionnaire is accurate, when any items are added or removed, these properties of the scale need to be checked again to ensure they have not been negatively affected [11]. Thus, adding to such a scale requires psychometric expertise to conduct these tests [11]. Since most people do not have this expertise, it would mean were the Genet tool based on this approach, it would be inextensible. While it is not impossible to automate the processing needed to run validity and reliability checks (see for instance [12]), this would require substantial effort.

2.2.1 *Summary table and approach selection*

Table 1 below summarizes the evaluation of the methods reviewed so far. In this table the "Y" signifies that the condition in the column header has been met, while "N" signifies that it has not. A "Y" in the "other" column signifies that the method has issues other than the four desiderata explicitly named, while an "N" signifies it does not.

Table 1 Evaluation of elicitation in techniques machine ethics

	Automatability	Plurality	Extensibility	Mapping	Other
Philosophical inventory	Y	Y	N	N	N
Moral Machine experiment	Y	Y	Y	N	N

As can be seen from the table, both approaches reviewed have numerous drawbacks as the basis for a Genet theory extraction tool. The most significant of these in both cases, as detailed above, are the mapping issues. Given this, neither of these approaches will be used as the basis for a Genet extraction tool prototype.

2.3 Elicitation in the cognitive sciences of morality

While there is little research in machine ethics on the elicitation of lay-people's ethical theories, there are many studies of exactly this in the cognitive sciences of morality (e.g., moral psychology, and experimental moral philosophy) [16].

Since the 1990s there has been significant overlap in moral psychology/cognitive science between psychology, and fields like philosophy and neuroscience [17]. In this interdisciplinary period, there have been two dominant elicitation techniques, vignettes based on thought experiments developed in the moral philosophical literature and psychometric scales [18]. While both are seen as equally reliable and valid, recently researchers have been moving toward the use of questionnaires/scales because one question is usually much shorter than one vignette/thought experiment, so questionnaires are the more efficient elicitation method [11].

A preliminary literature review has revealed that there are several psychometric scales capable of measuring the degree to which one holds one of the four ethical theories used in [1] to illustrate how Genet would model various ethical theories. For example: the Oxford Utilitarianism Scale is available for identifying one's level of utilitarianism (which can also be used to identify Kantian deontologists) [11]. The Morality Founded on Divine Authority (MFDA) scale [14] [18] for identifying one's level of commitment divine command theory (DCT). The Egoism scale for identifying one level of egoism also exists 16]. In some cases, there was more than one scale available to test for a particular moral theory (e.g., [20] and [11]).

In addition, there are also critiques of many of these scales. For instance, recent work [21] critiques the OUS as developed in [11]. These critiques focus on the core idea in [11], namely that utilitarianism is a composite of two components, and thus that the OUS must be a composite of two subscales. These subscales measure what they call "impartial beneficence" (OUS-IB) and "instrumental harm" (OUS-IH) [11]. Impartial beneficence occurs when a person acts in the best interests of another, no matter who that person is, if doing so would lead to the greatest good for the greatest number [11]. Instrumental harm occurs when a person is willing to undergo, or cause, harm to generate the greatest good for the greatest number [11]. The study in [21] provides both conceptual and empirical reasons for the idea that sacrificial dilemmas (and the questions based on them in one of the OUS subscales) do predict a tendency toward favouring impartial beneficence. There is thus no need to have a subscale which measures impartial beneficence and a separate one which measures instrumental harm, as the OUS does [21]. Despite this contestation, each of these scales have been empirically tested for reliability and validity (see for example [11]).

Given that these scales use the same categories as Genet does, there is no issue mapping between them. However, like other scale-based approaches, these scales are easily automatable but not as easily extensible (see for instance my discussion of [12] above). While extensibility is also still an issue, scales for a diverse set of theories already exist (e.g., [22]). However, thesescales, as well as the vignette-based presentation of moral dilemmas, have yet to be evaluated from a user experience or usability perspective [23].

A final issue that, while not captured by the set of criteria used in this review, is nonetheless important given the overarching aim of classifying users based on their preferred moral theory, is the model of lay moral cognition that has emerged from findings like [11] (but also the otherscale developers mentioned above). Here the focus will be on the position put forward in [11] since it explicitly discusses how explicit moral theory differs from what is observed in lay moral cognition [11].

To evaluate the validity of the OUS scale developed in [11], the authors used a number of validation measures that had independent support as measures of utilitarianism [11]. This included observing how answers on the OUS correlated with Likert scale responses of agreement with a description of the utilitarian moral theory in simple language [11]. They also employ the kinds of moral dilemmas presented as vignettes already discussed above as validation measures as well. In this study, n=960.

According to [11], the results suggest that people do not have explicit moral theories, but they

have tendencies the outputs of which (e.g., behaviours and judgments) can approximate the outputs of moral theories. According to them:

> It bears emphasizing again that the construct we have in mind is a measure of broad tendencies in moral deliberation and judgment in the lay population. It is not likely that ordinary people apply anything resembling an explicit ethical theory (whether utilitarian or not), nor is it likely that their moral judgments are fully consistent across different moral contexts. [11, p. 136]

If this view is right, then one would expect to observe what they did: small but significant correlations between OUS scores and explicit statements describing utilitarianism in layman's terms and between these scores and the validation dilemmas [11] (i.e., $r<=40$, $p<0.01$).

These findings establish the possibility that any project, such as the one undertaken here, to find a single preferred moral theory underlying lay moral cognition is unlikely to succeed, because there is none to be found. While this does not give clear guidance in terms of moral elicitation system design, it does establish a possible obstacle the present study may face.

2.3.1 *Summary table and approach selection*

Table 2 below lays out how each of the elicitation techniques reviewed in Section 2.3. compare in terms of the desiderata.

Table 2 Evaluation of elicitation techniques in the cognitive sciences of morality

	Automatability	Plurality	Extensibility	Mapping	Other
vignette/thought experiment	Y	Y	N	Y	Y
scales/ questionnaires	Y	Y	N	Y	Y

Given the above, the questionnaire/scale-based approach elicitation method will be used to represent the cognitive scientific approach. This is because, although methods perform equally well in terms of the desiderata, the scales are shorter as they are based on short questions rather than longer vignettes.

2.4 Elicitation in HCI

For our purposes, two areas of work in human-computer interaction (HCI) that are potentially applicable to designing a theory extraction tool for Genet: work on preference elicitation and in the "Value Sensitive Design" (VSD) [24] tradition. Much of the work on user preference elicitation in HCI outside the VSD framework is older work (pre-2010) that is mainly motivated by the need for elicitation of user preferences for new users of recommender systems [25]. The issue there is that new users will not yet have made selections on a given system, thus the system has no data it can use to infer the user's preferences, and thus it has no basis for good recommendations. This is known as the "cold start" problem [25]. This work will not be the focus here. The work that will be the focus is that of the VSD researchers and HCI

researchers from outside the VSD community who have nonetheless developed tools for eliciting moral information from users (e.g., [27]). This is because these morally focused elicitation tools are closer in focus to the moral theory elicitation task that the Genet theory extraction tool will have to complete.

2.4.1 *Value Sensitive Design*

VSD is an approach to the design of technological systems that foregrounds the interaction of these systems with the values of the stakeholders in contact with and affected by the systems [24]. It is thus an approach which tries to minimize the morally undesirable unforeseen consequences that may arise from this interaction between technology and human values if these values are not considered in the initial design of said technology [24]. To achieve this, proponents of VSD advocate for a "tripartite methodology" [24, p. 1], involving conceptual, empirical and technical inquiry. Here our focus will be on the empirical work, as it is the most concerned with eliciting values from actual end-users. However, given the nature of the current project, the conceptual (in considering the possibility of mapping from values to moral theories) and technical branches of the VSD approach will also need to be addressed (if these techniques are implemented computationally)[24].

Friedman, Hendry and Boring [24] provide a survey of the empirical methodologies employed in VSD. They review 14 such methodologies, labelling each according to the function they are usually are put to in the research process [24]. Since the main aim of the project being undertaken here is the creation of an elicitation tool, the focus will be on the subset of the methodologies reviewed in [24] that are labelled as useful for "value elicitation" [24, p. 3]. This reduces the total number of methods to be considered for our purposes to 6 [24]. These are: Value Scenarios, Value Sketch, Value-oriented Semi structured Interview, Scalable Information Dimensions; Value-oriented Mockup, Prototype or Field Deployment and the Value Sensitive Action-Reflection Model [24].

2.4.1.1 Value Sensitive Action-Reflection Model and Value-oriented Mockup, Prototype or Field Deployment

The Value Sensitive Action-Reflection Model is a technique for incorporating value prompts into co-design projects [24] and thus presupposes a co-design methodology. Value-oriented Mockup, Prototype or Field Deployment is also a co-design approach [24]. What differentiates it from other co-design approaches is its emphasis on the user's/co-designer's values as they are expressed throughout the design process [24].

Since both of these are co-design methodologies, they would be difficult to apply to the design of an elicitation tool for Genet. This is because, given the problem the designing an extraction tool for Genet presents, it makes it difficult to incorporate those without the relevant expertise into the design of such a system. For instance, it is unclear how, if users are unfamiliar with the ethical vocabulary Genet uses, they would be able to conceive of a way to extract this information from others.

2.4.1.2 Scalable Information Dimensions

This method entails creating sets of questions about a technology to be assessed/developed [24]. The questions focus on the range of levels along dimensions such as pervasiveness (e.g.,

"what if the user base grows to 1 in 10 of the world's population?", "What if it grows to 5 in 10?" [27, p.10]) and/or granularity of information (e.g., "What if the application requires your region of residence?", "What if it in order to see how much each factor/dimension influences a user's opinion/perception of the technology requires your street name" etc).

Since these methods just consist of asking sets of questions, it should be easily automatable. It should also be able to extract many different theories of morality depending on the questions selected.

The Scalable Information Dimensions technique also faces the same extensibility issues as the questionnaires and the other methods reviewed thus far. Since it would be easy to add a question but difficult to check its effect on accurate categorization.

Finally, whether the output of this method can be mapped onto theories that Genet can represent depends on the questions selected. Unfortunately, the method gives no guidance on doing this for eliciting moral philosophical theories.

2.4.1.3 Value-oriented Semi structured Interview and Value Sketch

This approach is identical to semi-structured interviews in the context of non-VSD research. But in the context of VSD, the interview tries to elicit more value-related information [24].

Since the questions in a semi-structured interview are not known beforehand, this approach is by definition difficult to automate.

It is also ill-suited to eliciting information that is not consciously accessible, since the straight-forward questioning is unable to elicit such information. Thus, it is ill-suited to the elicitation of moral theory, multiple or otherwise, from those who do not consciously have them.

Turning to the Value Sketch method, this is an umbrella term for approaches employed by VSD researchers which use sketches and/or sketching activities to elicit information about users' values [24].

These approaches face the same issue that faced the moral machine project's pictorial elicitation mechanism, which is the difficulty of pictorially representing abstract concepts.

2.4.1.4 Value Scenarios

Nathan et al. [28] developed the idea of value scenarios (VSs) within a theoretical framework which is a fusion of value sensitive design and "technological appropriation" [28, p. 9]. This perspective foregrounds the idea that when one introduces a technological factor into an environment (which can be thought of as a set of systems which predates the technological addition), it is likely to cause effects which are difficult to predict [28]. The VSs are an aid to envisioning and thus anticipating at least some of the morally salient effects that might have remained unforeseen [28].

While the approach is as automatable as the other text-based approaches, such as the psychometric scales, one issue this approach faces is that it is not clear how one would map from judgements in response to scenarios to the moral theory the user holds. It is true that the mapping depends on the content of the scenarios designed/selected, but it is a non-trivial task to design this content in a way that would allow one to use responses to it to infer the responder's moral theory. This would need to be done while also not simply copying the moral dilemma approach used in the vignette-based methods discussed in Section 2.3. Finally, it too

faces the extensibility concerns facing most of the other methods reviewed thus far.

Another relevant value-based (but from outside VSD) elicitation instrument is the Moral Foundations Questionnaire (MFQ) [29]. The MFQ is a scale which focuses on values rather than moral philosophical theories, which is more similar to the value-sensitive design approach than the scales reviewed in Section 2.3. Since there have already been thought experiments developed based on the MFQ [30], it should also be trivial to translate these into VSs (if that is the approach that is chosen). There has also been work such as [31] on using values measured by the MFQ to predict the level of belief in traditional moral theories such as utilitarianism, which would help with mapping the MFQ-based VSs to the moral theories discussed in [31].

2.4.2 *Preference elicitation in HCI*

Although most relevant work on elicitation methods for moral information can be found in the VSD literature, as mentioned at the start of this section, there are some exceptions. An example of such an exception will now be discussed.

Niforatos et al. [26] investigated the use of virtual reality to elicit moral judgement. They observed the judgements of participants (n=60) who responded to moral dilemmas [26]. One of the dilemmas used was the Trolley Problem [32]. This dilemma has one decide between allowing a train to kill multiple people, or diverting the train so that it only kills one. Unlike in most moral psychological studies where participants just imagine this scenario (and/or others like it) (see, for instance, Section 2.3), [26] simulated these experiences for their participants using virtual reality (VR) [26]. They observed that participants' judgements differed to responses in non-VR based versions of the same dilemmas [26].

Since this is a VR based approach, it is obviously automatable. Given that the approach is also thought experiment/dilemma based, it can represent options agreement/disagreement with which constitute an agreement/disagreement with some ethical theory. Since this is often the method used by philosophers to illustrate their views (see [33]), it should be able to present options which correspond to the moral philosophical dimensions Genet can represent. It is an under explored question however, whether VR is closer to pictorial representation (low) or textual representation (high) in the ease with which one can express an abstract concept.

This method would suffer from more serious obstacles to extensibility than the other methods reviewed thus far. Not only, as with other dilemma/thought experiment approaches, will validating additional items be difficult, but adding the difficulty of designing virtual reality experience would compound this.

2.4.3 *Summary table and approach selection*

Table 3 below lays out how each of the elicitation techniques reviewed in Section 2.4 compare in terms of the desiderata.

Table 3 Evaluation techniques of elicitation in HCI

	Automatability	Extensibility	Plurality	Mapping	Other
Value Sensitive Action-Reflection Model	Y	Y	Unclear	Unclear	Y

Value-oriented Mockup, Prototype or Field Deployment	Y	Y	Unclear	Unclear	Y
Scalable Information Dimensions	Y	Y	Unclear	Unclear	N
Value-oriented Semi structured Interview	N	Y	Unclear	Unclear	Y
Value Sketch	Y	N	Unclear	Unclear	Y
Value Scenarios	Y	Y	Unclear	Unclear	N
VR based approach	Y	Y	Unclear	Y	Y

Judging from the above, VSs and Scalable Information Dimensions perform equally well. However, while the mapping remains an issue for both, with VSs there are more options available to overcome this. This is because VSs are similar to the TEs, and there are MFQ-based TEs [30], as well as work of correlating MFQ value score with degree of belief in some moral theories [31].

2.5 Knowledge elicitation

Knowledge elicitation is the practice of extracting the epistemic content of mental representations from users so it can be implemented in a (knowledge-based) computational system [34]. The design of such a system falls within the field of knowledge engineering [34]. While expert systems used to be one of the more popular kinds of knowledge-based systems [34], since the turn of this century, this popularity has been usurped by ontology engineering [35]. With this shift of attention in the field, there has also been a shift in who produces research on knowledge elicitation [34]. Where research on knowledge elicitation within the knowledge engineering community used to be ensconced in the expert systems community, it has now moved to the ontology engineering community (sometimes under the banner of "Ontology elicitation" [34]). Finally, it should be noted that while, due to space and time constraints, approaches from human factors and requirements engineering which iterate on these elicitation techniques will not be discussed, these approaches do exist [34].

This section will survey methods from the knowledge and ontology elicitation research streams and their suitability as an approach to moral theory elicitation.

Special attention will be paid to indirect knowledge elicitation methods, as the issue facing the Genet extraction tool is that the moral theory of the user will in some cases, not be consciously accessible to said user (or at least will not be labelled using the same terms as those used by Genet). The method that this system will employ must be able to access representations not accessible to the user themselves. This is what indirect elicitation methods are designed for [36].

2.5.1 *Expert systems*

Hudlicka [36] compares three indirect knowledge elicitation techniques, Repertory grid analysis, Proximity scaling (including multi-dimensional scaling (MDS) techniques), and Hierarchical clustering analysis. Here only the first two of these techniques will be discussed, as they are the only ones which involve distinct approaches to eliciting information from the

user/expert (rather than merely a distinct approach to the analysis of this information once already acquired). Sections 2.5.1.1 and 2.5.1.2 thus both rely on the overviews provided in [36], unless otherwise stated.

2.5.1.1 Repertory grid analysis

The repertory grid analysis method was first formulated in the context of personal construct theory. This is a psychological theory developed by George Kelly [37]. This theory holds that humans represent the world as a set of entities (usually called objects or items of interest). These entities are distinguished by the set of attributes/properties they possess. These attributes are usually referred to as "constructs" [p. 17] by proponents of personal construct theory. According to this theory, behaviour (both mental and physical) is a result of the carrying out mental operations (both consciously and unconsciously) on these internal representations of constructs.

Repertory grid analysis is named after the repertory grid, which is the structure used to represent the objects the experts organize their domain of expertise into, and the constructs these objects possess. This is a 2-dimensional grid which represents the objects and groupsthem according to their constructs.

Repertory grid analysis is a three-step process. The first step of the process is to select objects/items from the domain of interest. The second step is to compare the selected itemsby asking experts in the domain of interest to list any similarities and differences between them. The final step is to ask the experts to rate each item along each of the attributes identified. This rating is then captured as a value on the appropriate cell on the repertory grid.

The repertory grid can then be used directly to help answer questions about the nature of the mental representations of the experts who help produced it or it can be further analysed. Itcan also be used as input into other indirect elicitation techniques.

Although the repertory grid method has decreased in popularity along with expert systems, and there has long been work on improvements (e.g., [38]), it is a technique that is still sometimes used. For our purposes here the most relevant (relatively) recent application of the technique was done by Lee et al [15], who used it as part of a system designed to elicit information about the mental representations humans use when engaging in moral reasoning and, in turn, allow the system to reason similarly.

Since Lee et al [15] show that the approach can be computerized, it is clear that the approach is automatable. Some guidance is also available for the initial selection of items of interest, so there is some help available for gathering the content that will be used by the Genet extraction tool [15]. However, extensibility is still an issue.

For this method, the issue of mapping can be dealt with in the item selection stage. For instance, Lee et al [15] recommend going directly to the philosophical literature to search for thought experiments that can be implemented as items.

There is one large issue with the repertory grid analysis approach. The issue is that it requires an expert to assist with the construction of the grid once the data has been collected from the study's participants. This limits its automatability.

2.5.1.2 Proximity scaling

Proximity scaling techniques are a set of methods designed to indirectly access the content of mental representations via judgements about the similarity of objects in the domain of interest. This method specializes in the investigation of the mental representations of complex perceptual input.

The proximity scaling methodology requires two steps. The first step is to elicit the proximity (i.e., similarity/dissimilarity) judgements. These judgments are prompted with pairs of objects from the domain of interest as the stimulus. How this is done varies depending on the objectives of the researchers. This information is then stored in a matrix,known as a proximity matrix, where each cell represents the similar/proximity between two domain elements.

This proximity matrix is then used as input to an algorithm (which, again, differs depending on the project) that outputs a representation of the global mental representation (e.g., a network, or map). In many cases, this global representation is then shown to experts for further information (e.g., the meaning of links a network/ graph).

While proximity scaling techniques are similar to the repertory grid analysis method, they do differ in a few important respects. Firstly, while repertory grid analysis can ask participants for similarity judgements of more than two entities/constructs in a domain, proximity scaling techniques usually only elicit similarity judgements between pairs of items from a domain. Secondly, they differ in their outputs. Proximity scaling is a spatial representation where items are positioned in such a way that the distances between them reflect the original proximities (i.e. similarity judgements). This helps in identifying clusters, dimensions, and the underlying structure of the data. Repertory Grid Analysis on the other hand, produces a grid (matrix) showing the relationship between elements and constructs where the main information is not encoded by positions of the items relative to each other, but by the actual values given per construct/attribute. This grid can then be analyzed to understand the individual's cognitive structure and the dimensions they use to differentiate between elements.

Verheyen and Petersen [39] demonstrate that proximity scaling techniques can be used in the moral domain. They used an elicitation technique whereby they selected 10 TEs and compiled a list of five moral principles. Participants were then asked to make pairwise comparisons of each of the 10 thought experiments (yielding 45 comparisons in total) in response to the prompt asking how morally similar each pair of cases is. After each of these comparison judgements, the participant was asked which of the five principles on the list of principles compiled by the experimenters they would endorse as a guide for action in each TE. Multidimensional scaling (MDS) techniques were used to construct a similarity space representing the range of cases covered by each principle.

Like the repertory grid approach, proximity scaling is automatable as it has been automated in other areas such as social psychology [40]. Proximity scaling also resolves the mapping issue in the selection of the domain elements to compare.

One issue is that the proximity scaling method was designed with an emphasis on accessing perceptual representations [36]. The representations which encode moral beliefs are not perceptual [16]. As [39] demonstrates, this does not preclude proximity scaling techniques from being applied to the representations underlying moral judgement.

Another issue is that Genet requires output from the elicitation system in terms of a determinate categorization of the user into one or another disjunctive category based on their preferred moral theory. However, the similarity space the method in [39] outputs does not provide this.

It may be possible to fix this by calculating the size of the regions covered by each principle, and assuming that the principle that covers the largest area in that space should be taken as reflective of the that user's moral theory. Furthermore, with proximity scaling techniques extensibility issues occur.

2.5.2 *Ontology engineering*

Within the ontology engineering literature, "bottom up" [34, p. 139] methodologies are most applicable. This is because they do not rely on existing ontologies and thus require one to construct a new ontology based on information acquired from relevant sources, one of which is often the relevant experts [34]. So, some of the techniques these researchers use will thus be able to extract information from human beings, much as the Genet extraction tool is supposed to.

The literature search conducted for the present review has shown that new research in "bottom-up" elicitation of information from humans has either not automated (e.g., [41]) or not concerned with automating the knowledge acquisition process for implicit knowledge (e.g.,[42]). Research that tackles the elicitation of implicit knowledge often recommends observation methods (e.g.,[43]).

However, these observation approaches without a theory of the domain of interest to guide them are blind. This is because without a model of what one is looking for, it is not clear which parts of a phenomenon deserve attention. Since these recommended observation methods such as contextual inquiry have yet to be used in the context of moral theory elicitation, no guidance exists for what one ought to observe to infer a person's moral theory [44]. Thus, this method is ill-suited to do the extraction task that is required by Genet.

Other relevant work from the ontology engineering community is that of Costechi [45][46]. He describes the problem he addresses similarly to that facing the development of a genet theory extraction tool [45]. He sees the problem as that of developing a way for a computational system to interact with a user such that it can extract the content of their relevant mental representations [45]. But while Genet's extraction tool must access mental representations encoding the user's preferred ethical theory, [45] is concerned with the mental representations which encode ontological knowledge. As a solution, [45] [46] attempts to demonstrate the soundness of using a dialogue system to automate the knowledge/ontology elicitation interview.

The drawback of this approach is that it lacks a way of determining the content of the interview questions used in an interview to elicit moral information. Since this is not a trivial thing to produce, adopting such an approach would require significant extra work that the approach itself is unable to guide.

2.5.3 *Summary table and approach selection*

Table 4 below lays out how each of the elicitation techniques reviewed in Section 2.5 compare in terms of the desiderata.

Table 4 Evaluation of elicitation techniques from requirements and knowledge engineering

	Automatability	Plurality	Extensibility	Mapping	Other
Repertory grid analysis	Y	Y	Unclear	Y	N
Proximity scaling	Y	Y	Unclear	Y	N
Observation	Unclear	Unclear	Unclear	Unclear	Y
Costechi dialogue system	Y	Unclear	Unclear	N	N

As can be seen from the table above, the two most suitable approaches for the Genet extraction tool are proximity scaling and repertory analysis. The proximity scaling approach will be employed, as it has been shown in [39] to apply to the moral domain, and is the only one of the two which does not rely on expert judgement, which would complicate the automatability of the technique.

3. DESIGN OF PROTOTYPES

This chapter starts by discussing which approaches reviewed in Chapter 2 will be selected for investigation, modification and implementation as prototypes in this project. The design of the moral theory categorization algorithms used in each prototype will then be discussed. Next is a description of the architecture and component structure that all the prototypes share. The differences in the process flow between CogSci-PT and HCI-PT on one hand and KE-PT on the other will then be discussed. Next, the design of the UI of each prototype, in terms of the moral theory elicitation stimuli and interaction sequence used, will be discussed. The technology used for each prototype will then be described. Finally, the key differences between the prototypes will be discussed.

3.1 Preliminaries

The three disciplines selected for investigation are cognitive science, and more specifically moral psychology; HCI, and more specifically VSD; and finally, knowledge engineering, and more specifically knowledge elicitation. The specific approaches that have been selected from each discipline are: moral psychological questions and statements representing cognitive science; VSs, representing HCI/VSD; and TEs, representing knowledge engineering/ knowledge elicitation.

The approaches above are selected because they best meet the criteria of facilitating the kind of mapping needed, having the necessary plurality, automatability and extensibility; as discussed in Sections 2.3.1, 2.4.3 and 2.5.3.

Each of these approaches has different characteristics, so each needs to get their own prototype with a way of interrogating the user and computing the answer. The prototype representing the cognitive science approach will be called “CogSci-PT” (CogSci-PT) and will rely on moral psychological questions and statements (as discussed in Section 2.3) as its elicitation stimuli.

These questions and statements are embedded in moral psychological scales. The scales that will be used in the first prototype are the OUS [11] for measuring a user’s tendency toward utilitarianism. This scale will also be used to measure the user’s tendency toward deontology. The MFDA [18] will be used to measure a user’s commitment to DCT. Finally, the ES [19] will be used to measure a user’s tendency toward egoism. The OUS scale has nine questions and requires responses on a 1-to-7-point Likert scale. The MFDA scale has five questions and requires responses on a 1-to-9-point Likert scale. The ES has 20 questions and requires responses on a 1-to-5-point Likert scale. Each scale length here excludes attention checks, and in each case the maximum Likert response signifies maximal agreement. Together, these scales amount to 34 questions.

The prototype representing the HCI approach will be called “HCI-PT” (HCI-PT) and will implement VSs (as discussed in Section 2.4.1) as its elicitation stimuli. HCI-PT is based on vignettes themselves based on the MFQ, which are modified by the author of this thesis to approximate VSs. The prototype will present users with VSs closely based on the vignette-based version of the MFQ designed by Clifford et al [30]. Six values are of interest for our purposes, these are: care, sanctity, loyalty, authority, fairness and liberty. All the values besides care are measured by five scenarios. While the care value is measured by six questions, three for care for animals, and three for care for other humans. This will result in 31 questions in

total (excluding attention checks). All scenarios will require a response on a 1-to-5-point Likert scale.

Finally, the prototype that represents the knowledge engineering approach will be called "prototype 3" (KE-PT) and it uses TEs (as discussed in Section 2.5.1) as elicitation stimuli. Two TEs were used per moral theory discussed in [1], resulting in eight TEs in total (excluding attention checks).

All the scales, VSs and TEs used can be found in Appendix A.

3.2 Algorithm design for theory classification for each prototype

For each approach, an algorithm was designed so that, based on the answers given by a user in response to the elicitation stimuli, it will compute the theory they adhere to according to those answers. The details of the algorithm design for each of these prototypes are discussed in the sub-sections below.

3.2.1.1 Design and description of classification algorithm for CogSci-PT

The first part of the algorithm for CogSci-PT implements the same scoring procedure that the authors of the scales use (i.e., averaging the Likert responses to each scale). Specifically, the prototype sums raw ratings on each scale or deducts the rating if question is negatively scored. The negative scoring occurs where stronger agreement on the Likert scale does not mean a stronger belief in the moral theory being measured. For instance, in the MFDA, the statement "It is possible to live a righteous life without knowledge of God's laws", is negatively scored, since strong agreement suggests a low credence in DCT. The mean score on each scale is then calculated.

At this stage in processing, the user has thus been assigned an OUS score, an ES score and an MFDA score (which are averages of the responses to each scale). All the steps thus far amount to following the same procedure as the scale designers for scoring the scales.

Since the process above outputs a score set and not a single score indicating that a participant's preferred theory is, further processing is necessary. This further processing entails comparing the scores initially outputted. To do this, the prototype converts the mean scores on each scale into a percentage of the maximum score on each scale. For instance, for the OUS scale, since the responses are on a Likert scale of 1 to 7, with 7 being maximal agreement, the OUS percentage = (OUS mean/7)*100. The percentage conversion is necessary because each scale employs Likert scales of different lengths, making direct comparison of the means derived from them untenable. The algorithm then compares these percentages to reach a classification. These comparisons are done according to five rules. Both the conversion to an average score, and the rules for classification were the contributions of the author of this thesis (with some caveats that are clarified in the discussion below).

These rules can be seen in the part of the algorithm reproduced in Figure 1 below:

```
1.  if percentage_ous<50 and percentage_mfda<50 then
2.      if percentage_es<50:
3.          return="You are a deontologist"
4.      endif
5.  else if percentage_ous> percentage_mfda and percentage_ous>percentage_es then
6.      if percentage_ous>50:
7.          return="You are a utilitarian"
8.      endif
9.  else if percentage_ous<percentage_mfda and percentage_mfda>percentage_es then
10.     if percentage_mfda>50:
11.         return= "You are a divine command theorist"
12.     endif
13. else if percentage_es> percentage_mfda and percentage_es>percentage_ous then
14.     if percentage_es>=25:
15.         return= "You are an egoist"
16.     endif
17. else if percentage_es==percentage_mfda and percentage_ous== percentage_es then
18.     if percentage_ous>50:
19.         return="Your moral theory is not ascertainable by Genet at this time"
20.     endif
21. else if:
22.     return="Your moral theory is not ascertainable by Genet at this time"
23. Endif
```

Figure 1. Section of the algorithm implementing the rules for theory categorization for CogSci-PT

The first two rules on lines 1-2 and 5-6 in the figure above determine if the user is a utilitarian or deontologist. These rules are adapted from a rule used in [11] to sort utilitarians from deontologists. This rule categorized those who scored above the midpoint as utilitarians, and the rest as deontologists. However, this version of the rule did not also incorporate a check of other moral theory scores. These additions were designed by the author of this thesis by generalizing this midpoint threshold idea to the DCT categorizations as well by using the MFDA percentage>50 as indicative of belief in DCT. Finally, since the mean for the ES scale reported in [19] is lower than the scales', ES percentage>25 was used, instead of the higher threshold used for the other scales. This egoism rule was also designed by the author of this thesis.

For the purposes of the prototype, the two subscales of the OUS were collapsed. This design choice was informed by the critiques in [21].

3.2.1.2 Design and Description of classification algorithm for HCI-PT

The first part of the algorithm implemented in HCI-PT adds the Likert rating given in response to each VS to cumulative score for each value measured (i.e., care, fairness, loyalty, sanctity

and authority). For the scenarios that were negatively scored, the rating given is deducted from the relevant value score. These cumulative scores are then averaged, resulting in an average score for each value. This is the same scoring procedure used in [31].

For the mapping between these value scores and the utilitarianism and deontology classification, a dilemma score was used. The dilemma scores were calculated following the procedure used by Crone and Laham [31], which uses a multiple linear regression equation, which takes the value scores, and results reported in [31] including the regression weights and means of the dilemma scores and values scores as parameters. The dilemma score was the predicted independent variable obtained from this calculation, as follows:

$$D=3.25-0.21\times C-0.09\times F+0.04\times L-0.04\times \mathrm{A}-0.16\times \mathrm{S}$$

Where *D* represents the dilemma score, *C* represents the care score, *F* represents the fairness score, *L* represents the loyalty score, *A* represents the authority score, and *S* represents the sanctity score. The y-intercept value (3.25) was calculated using the follow equation [47, n.pag] (which was derived from the multiple linear regression equation):

$$b_0 = \bar{Y} - (B_1 \cdot \bar{X}_1 + B_2 \cdot \bar{X}_2 + \ldots + B_k \cdot \bar{X}_k)$$

Where $b0$ is the y-intercept, $\bar{Y}$ is the mean of the independent variable (i.e., the mean of the dilemma scores [31]), $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_k$ are the means of the dependent variables (i.e., the means of the MFQ values as reported in [31]), and ($B_1, B_2, \ldots, B_k$) are the regression weights (which are given directly as regression weights in [31]).

This approach was used for the dilemma score (which maps onto the utilitarian and deontology categories) but not the DCT predictors because what they are predicting in [18] are scores on the MFDA (which are already implemented directly in CogSci-PT). Thus, a DCT categorization rule based on a measure with less overlap (i.e., the value scores themselves) was implemented in HCI-PT.

Rules were designed by the author of this thesis and used to classify the users' according to their moral theory based on the value and dilemma scores. These rules were based on the average scores reported in [31] and [18] (plus or minus standard deviation). For the egoism categorization, hypothesized correlations were used based on conceptual relationships between egoism and the values measured by the MFQ. For instance, since the care value is supposed to capture care for others (and animals), it should be negatively correlated with egoism, which centers on one's own interest and wellbeing. These classification rules were designed by the author of this thesis.

These rules, expressed algorithmically, are included in Figure 2.

```
1. if loyalty_score>= 3.34 and authority_score>= 3.64 and sanctity_score>=
   3.65 and dilemma_score<=1.58 then
2.     return="You are a divine command theorist"
3. endif
4. else if loyalty_score<= 1.5 and care_score<=2.6 and fairness_score<=2.46
   and authority_score<= 1.84 and sanctity_score<= 1.15 then
5.     return="You are an egoist"
6. endif
7. else if dilemma_score>=1.58 then
8.     return= "You are a utilitarian"
9. endif
10. else if dilemma_score<1.58 then
11.     return="You are a deontologist"
12. endif
```

Figure 2. Segment of the algorithm implementing the rules for theory categorization for HCI-PT

The threshold for the deontology and utilitarian categories is 2.70 (average "sacrifice rating") minus the standard deviation (1.12) of [31]. This lower threshold was used because it was observed in testing that using the mean of [31] led to an implausibly low rate of utilitarianism classifications. The thresholds for the DCT (line 1 in the figure) classification is informed by [18], where the rules used for classification check the particular value scores that were shown to correlate with adherence to DCT [18]. The MFQ values reported in [18] as correlating positively with DCT belief (as measured by the MFDA) are the Authority, Sanctity, and Loyalty values. Since the raw or average MFQ scores are not reported in [18], the MFQ averages and standard deviations reported in [31] were used. The thresholds for loyalty (M = 2.42, SD = 0.92), authority (M = 2.74, SD = 2.74) and sanctity scores (M = 2.40, SD = 1.25) used for categorization were one SD above the mean (threshold score = M+SD). The same dilemma score as for the deontology and utilitarian categorization was used in the rule for DCT categorization as well. A lower threshold was used for the scores besides the dilemma score because it would ensure that those assigned to these categories are the most likely to be DCT proponents, since they would be in the upper range of DCT predictors.

The egoism categorization rule (line 4 in Figure 2) was informed by the fact that egoism is a kind of consequentialist moral theory, and thus would be negatively correlated with deontological thinking. This implies a higher-than-average dilemma score. It was also postulated that egoism would negatively correlate with the values measured by the MFQ, since none of these values capture the self-regard central to normative egoism. Thus, the thresholds used were the averages reported in [31] minus their standard deviation. So, in addition to the MFQ value averages and standard deviations discussed above, the care and fairness value mean and SD were used in the rule for calculating (threshold score = M-SD) the fairness and care thresholds in line 4 above (M = 3.46, SD = 0.86 and M = 3.33, SD = 0,87 respectively).

3.2.1.3 Design and Description of classification algorithm for KE-PT

Since Genet requires a determinate categorization of the user into one or another disjunctive category based on their preferred moral theory, the similarity space outputted by Veheyen and Peterson's method [39] as discussed in Section 2.5.1 is insufficient because it is not obvious how to infer the user's preferred moral theory from the similarity space produced by the proximity scaling technique. A better technique is to count the number of TEs the user applies each principle to, with the principle that the user applies to the most TEs being the users preferred moral theory. Yet, to do this, a similarity space is not necessary. Thus KE-PTdid not make use of the proximity scaling technique nor the multidimensional scaling computation.

Since there are six questions with four moral principles to choose from, it is possible to have a situation where two or three different moral theories are chosen with equal frequency. For these cases, a tie breaking procedure designed by the author of this thesis was added to the algorithm. This procedure finds which moral theories are tied by checking which principles have an equal frequency count, and stores the TEs in response to which each of the tied principles were given. At this point, two or three sets of TEs have been isolated (each corresponding to one of the tied principles). The algorithm then displays a question page to the user with one question for each of the tied theories. The question "which of these situations described above presents the most important moral decision?" is also displayed on this page. The algorithm then allows the user to select one of the TEs on the page in response. The algorithm then increments a count of the principle frequency for the tied principle that corresponds to the TE that was selected as most important. This loop repeats until all questions have been displayed. Once this is done, the principle with the highest frequency count (which represents the principle applied to what the user judges to be the most important situations) is the principle that represents the user's moral theory. If there is another tie, the process terminates with a page telling the user their moral theory could not be ascertained.

3.3 Software design and implementation of prototypes

In this section the software implementation of the prototypes will be described. First, the aspects of the system architecture and components shared by all the prototypes will be discussed. The interaction of the components of the prototypes will then be discussed. Next the UI implemented in each will be addressed. Finally, the technologies used to engineer the prototypes will be discussed.

3.3.1 *Shared architectural aspects*

Architecturally, each prototype has the same layered architecture, with a data-storage layer, a business logic layer and presentation layer. The data storage layer includes the shared database in which all the prototypes have two tables each. Each prototype has a "users" table in this database. This stores a user ID (as an automatically incremented integer), the questions users answered, the answers they gave, whether they were paying attention, and the classification they were given by the prototype they interacted with. They also have a "questions table", which stores the relevant TEs, scenarios or questions used by the prototypes.

The business logic layer handles the initialization of each prototype and implements the algorithms described in Section 3.2. This layer also interacts with the data storage layer to retrieve questions and store responses and classifications.

The presentation layer handles the display of the initial startup page, the pages with questions, and the results page.

The prototype components have a similar structure (especially CogSci-PT and HCI-PT). The data storage layer is implemented by a relational database. The business logic layer is implemented across the initializer, answer processor and results generator components. The presentation layer is implemented by the question asker and results generator components.

3.3.2 *Description of process flow implemented by the prototypes*

A process flow diagram, with more detail about how the components interact in CogSci-PT and HCI-PT, can be seen in Figure 3 below. KE-PT differs somewhat from the description in Figure 3, as it allowed for ties and needed components to deal with these. A diagram illustrating the process flow of the tie breaker system and the rest of KE-PT can be seen in Figure 4 below:

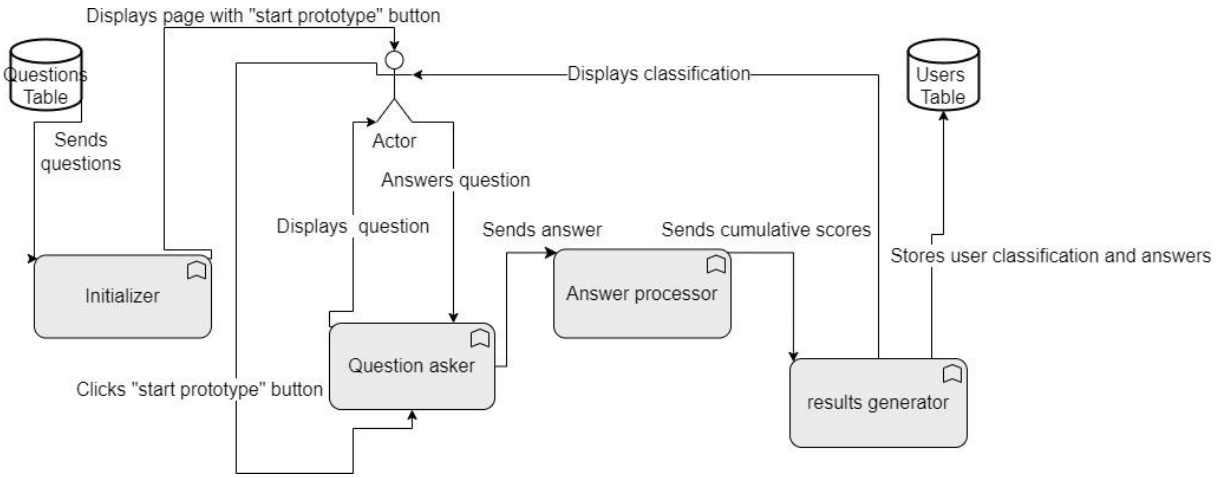

Figure 3. Process flow diagram representing the structure of CogSci-PT and HCI-PT

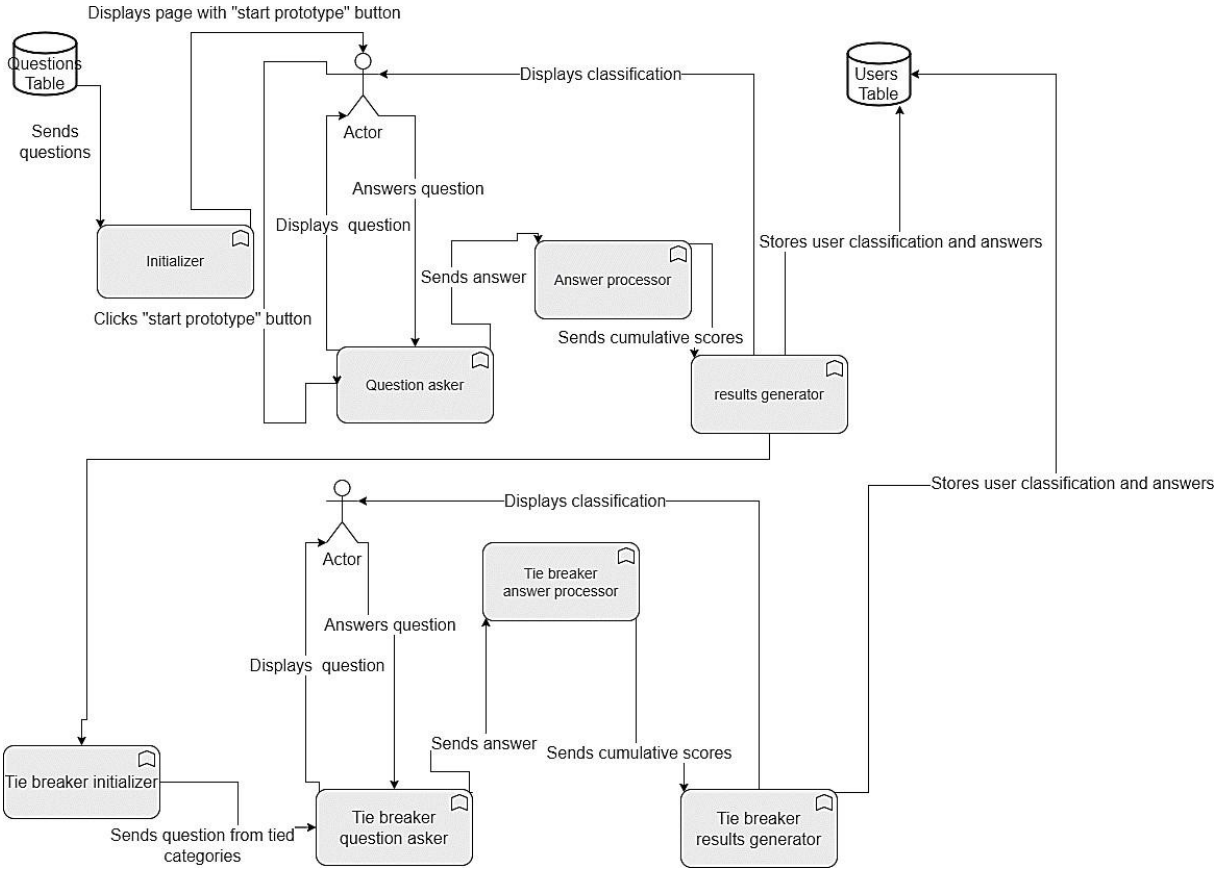

Figure 4. Process flow diagram representing the structure of KE-PT. Note that the two "actor" symbols denote the same user.

3.3.3 *Detailed description of the UI of the individual prototypes*

We now discuss detailed descriptions of the UI for each prototype, with particular attention paid to the elicitation stimuli used and the interaction sequences implemented.

3.3.3.1 Details of elicitation stimuli, and interaction sequence for CogSci-PT

From the user perspective, CogSci-PT sequentially displays each question or statement on each of the relevant scales and asks for a response (as a judgement on a Likert scale).

Once all the questions from all the scales have been displayed and responded to, the result is then displayed to the user. An example of the question and results pages displayed to the user by the prototype can be seen in Figures 5 and 6 below respectively.

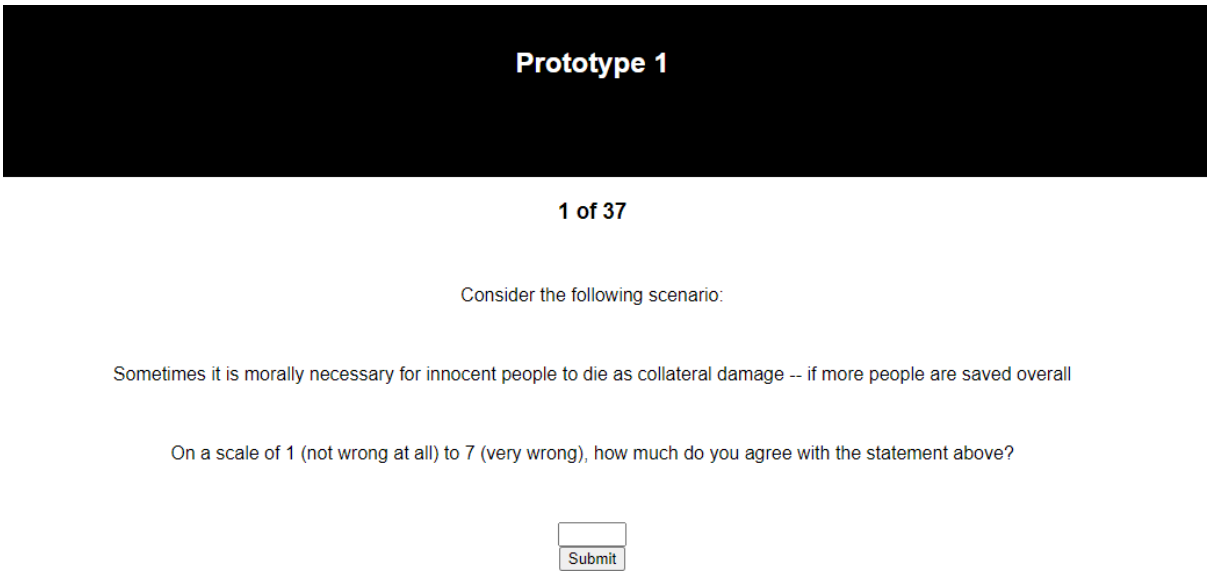

Figure 5. Screenshot of a question page in prototype CogSci-PT.

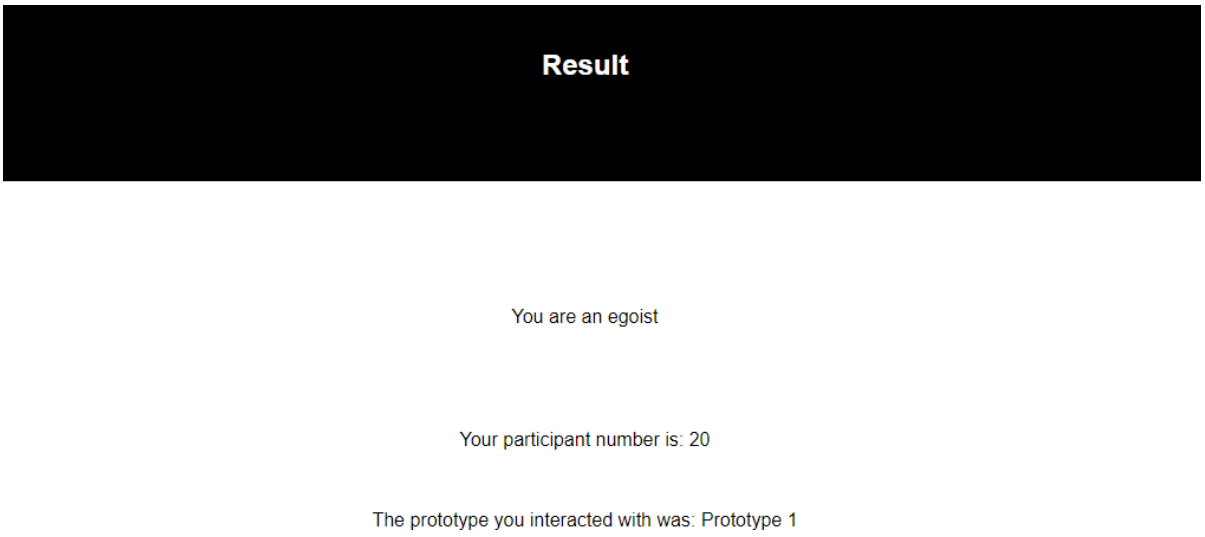

Figure 6. Screenshot of a results page in prototype CogSci-PT.

3.3.3.2 Details of elicitation stimuli, and interaction sequence for HCI-PT

The VSs used in HCI-PT emphasize four of the five elements of VSs as put forward in [28]. The emphasized elements are stakeholders, time, pervasiveness and value implications. The element not being emphasized is that of systemic effects. This is because the original MFQ vignettes emphasize concrete interactions usually between two participants, which makes it impossible to introduce content addressing systemic affects without completely changing (as opposed to just adapting) content of the original vignettes.

An example of the question page displayed by the prototype can be seen in Figure 7 below.

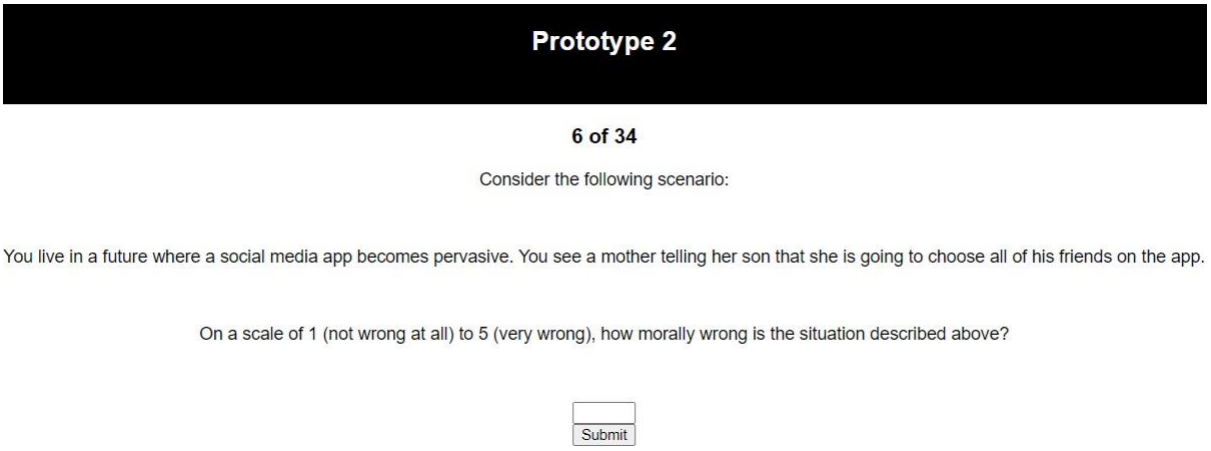

Figure 7. Screenshot of a question page in HCI-PT

3.3.3.3 Details of elicitation stimuli, and interaction sequence for KE-PT

KE-PT is based on the approach of [39]. What was taken from this work was the idea of using TEs and asking users to select one of a list of moral principles that they would apply to each TE. An example of the question page displayed to the user by the prototype can be seen in Figure 8 below.

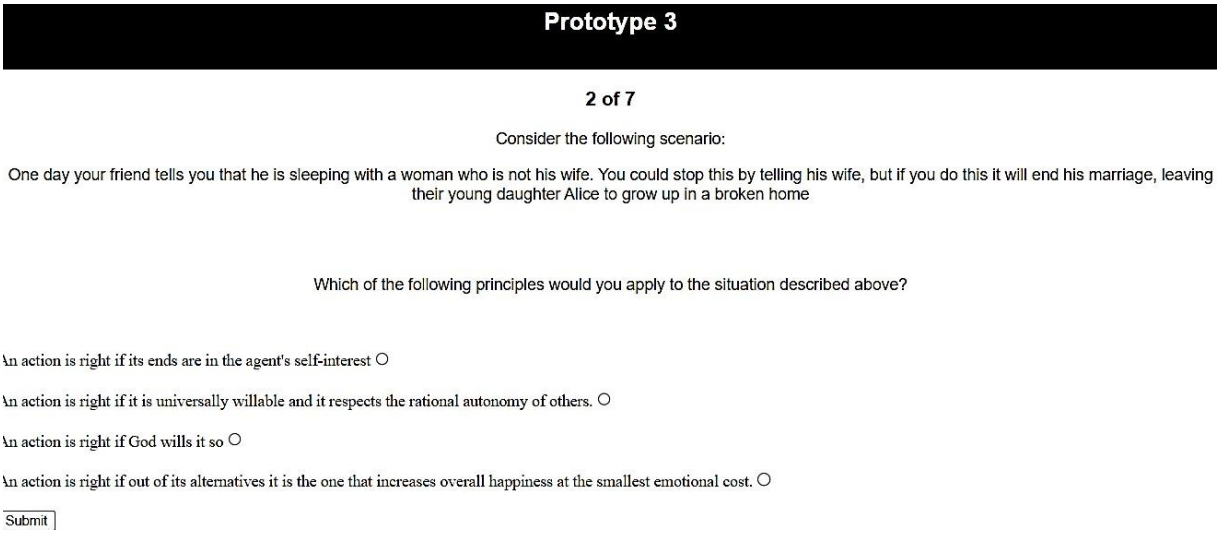

Figure 8. Screenshot of a question page in KE-PT

For the tie-break functionality, the user was presented with the question “Which one of the following two scenarios presents the most important moral decision?”, with two scenario descriptions below, and an adjacent radio button. An example of the tie breaker question page is in Figure 9 below.

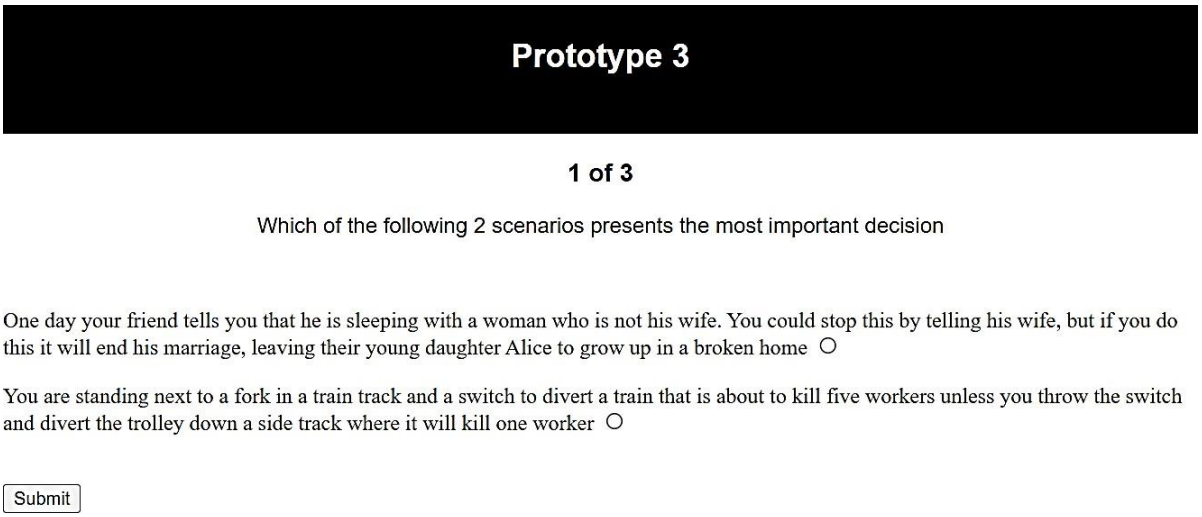

Figure 9. Screenshot of a sample Tie breaker question page of KE-PT that aims to realise a TE.

Once all the scenarios for each of the tied moral scenarios are compared, the result is then displayed to the user. If this leads to another tie, the user is served a webpage that informs them that their preferred moral theory could not be elicited.

For the prototype, the principles available to the user were those listed in [1]. The TEs used in the prototype were either drawn from the philosophical literature (e.g., [48] and [32]) or were specially designed for the prototype based on TEs in that literature.

3.3.3.4 Technologies used for all prototypes and other implementation details

Each prototype was built using Django 4.0.6, Python 3.8 and MySQL 8.0 for implementing the data storage and business logic layer. CSS 4 and HTML 5 served within Django views functioned as the implementation of the presentation layer. Each component described above is implemented by one Django view representing each of the components. The content of each prototype was randomized to avoid order effects. Each of the prototypes also presented a number of attention checks to determine if the participant is paying adequate attention during the interaction sequence. The attention checks take the form of prompts like “This is a control question to check whether you are paying attention. Please proceed by selecting utilitarianism on the scale below”. If all these checks are passed, the results page displayed to the user also contains a link to the URL for the validation tasks (which will be detailed in Section 4.3). If the checks fail, the result page is still displayed, but lacks this link.

Please note that the data collection tool actually deployed was a unified web application combining all three prototypes and included an informed consent page that preceded interaction with one of the prototypes (assuming consent was given). This was implemented within a single Django app. This application was hosted on an Apache server 2.4.29, which ran on a Linux an

Ubuntu 20.04.6 virtual machine provided by the UCT computer science department. The domain on which these prototypes were hosted was also provided by the UCT computer science department.

3.4 Summary of key differences between the prototypes

Table 5 below summarizes the key differences between the prototypes in terms of the elicitation stimuli and interaction sequence used; the values used as measures of the users' degree of belief in each moral theory, and the rules used for moral theory categorization.

Table 5. Differences between the prototypes

Prototype	CogSci-PT	HCI-PT	KE-PT
Elicitation stimuli and interaction sequence	Uses MPS/Qs as elicitation stimuli.	Uses VSs as elicitation stimuli.	Uses TEs as elicitation stimuli. And Takes principles as input.
Values used as measures of the users' degree of belief in each moral theory	Average score on each scale as a percentage.	Average score for each value as measured by the modified MFQ	Frequency count of number of times each principle was selected
Moral theory categorization rules	Uses midpoint of the scales (or the midpoint between the first and second quartile of the scales).	Uses mean of MPQ value scores +/- their standard deviation. And Calculates the dilemma score via multiple linear regression.	Simple frequency comparison of principle selection.

4. METHODOLOGY AND EVALUATION

This chapter will discuss the methodological and evaluative techniques that will be used to test the prototypes described in the previous chapter.

4.1 Aim(s) and participant recruitment strategy

The aim of the evaluation is to see which prototype performs best along the dimensions of validity/accuracy and usability.

Regarding the participant recruitment strategies employed, in this project, purposive sampling was used. This was done by posting announcements to the students in CSC1016 and CSC5014Z courses at UCT. A notice was also posted on the Google group named "Zaphil". This is a group for professional philosophers and postgraduate philosophy students in South Africa.

Ethics approval was obtained prior to the data collection stage from the UCT Faculty of Science Research Ethics Committee.

4.2 Outline of the data collection process

Participants interacted with one prototype which was randomly assigned from the pool of three prototypes. This randomization was done automatically once the participant visited www.elicitationtool.cs.uct.ac.za/Genet_prototypes/ and consented after having read the informed consent form.

The "task" the participants had to perform was to follow the interaction sequence they were guided through by each prototype. Once the interaction session with the prototype was completed and the categorization result was displayed, the participant was presented with a button that links to the validation survey hosted on LimeSurvey. This survey first asked several questions to obtain the relevant information from the participants regarding their philosophical expertise. The survey then had the participant complete the sequence of validation tasks which will be described in the next section. Participants were then asked to complete the System usability Scale (SUS)[49].

4.3 Validation

To ensure the approach selected and tool developed accurately extract users' moral beliefs, three kinds of validation methods were used. These were each modelled on the validation measures used in [11].

Each of these were implemented in a LimeSurvey survey. Once participants were classified by one of the three prototypes, they then interacted with each of these validation measures in sequence.

The first of the validation procedures did not require novel user input. Since there were professional philosophers in the sample, one way of testing whether the theory the participant was classified as holding is accurate is to see if moral philosophers, who know their moral theory in the terms native to Genet, report holding the same theory detected by the prototype. If there is such agreement, then this means they have been classified correctly, if not, then this is evidence that they were not correctly classified.

The second validation method is to check if participants' judgements in response to vignettes/scenarios describing ethical dilemmas that have not already been given to them align with the moral theory the prototype assigned to them. This was done by presenting the participants with these vignettes in the form of a section of the validation survey on LimeSurvey. These asked for response on a Likert scale of 1-5 (with 5 being maximal agreement and 1 being no agreement). More specifically, there were seven scenarios, six of which directly tested a single moral theory. Two covered DCT, two covered egoism, one covered utilitarianism and another covered deontology. Responses to each of these got added directly to the scenario score of the relevant moral theory (with negative scoring used where necessary). The remaining scenario counted either toward the deontology or utilitarianism score, depending on if the response was <3 or => 3. This resulted in the utilitarianism scenario score having a maximum of 9, the deontology and DCT scenario score having a maximum of 10, andthe egoism score having a maximum of 8. These scores are computed as percentages, since these totals are not equal.

The last validity measure took the form of a set of questions on the survey. This was a set of questions asking participants, on a Likert scale of 1-5, how much they agree with explicit ordinary language descriptions of each of the moral theories covered. If the users were accurately categorized, they would agree more with their prototype assigned moral theory than other theories.

For the usability testing part of the study the SUS was used. After participants had interacted with the prototypes and the other validity measures, participants were asked to complete the SUS.

4.4 Data Analysis

The procedure for data analysis described below was implemented as a set of Python 3.8 scripts which used the SciPy 1.11.1, Numpy 1.24.3, Pandas 2.0.3, Pingouin 0.5.3 and Statsmodels 0.14.0 packages to do the required computations.

First, three correlation values were obtained. These three values measured:

1. The degree to which the categorization given by the prototypes correlated with the level of agreement with the explicit statement.
2. The degree the categorization given the prototypes correlated with the answers to the vignettes/scenarios.
3. The degree to which the moral commitments of the philosophers in the sample correlated with how they have been classified by the prototypes.

For CogSci-PT and KE-PT these were all calculated as Pearson's Rs. For HCI-PT however, the DCT and Egoism categorizations were decided by rules that relied on the relationships between multiple MFQ value scores [19][18]. Thus, to check the accuracy of these categorizations using the relationship between the DCT and Egoism related value scores on the one hand, and the Egoism and DCT related validation scores on the other hand, Pearson correlation coefficients could not be computed due to there being a being multiple independent variables (the value scores). So, Pearson's Rs were computed between the prototype-assigned dilemma scores and the utilitarian and deontological validation scores only. The relationship between the HCI-PT scores and the validation DCT and egoism scores had to be computed using multiple linear regression. Thus, in these cases R,' the coefficient of multiple correlation, was derived by taking the square root of the R^2 statistic. In terms of the p-values for these, a

series of t-tests were used to produce the p value for each of the dependent variables in these cases.

Before the comparison procedure is detailed, there are a few things about the comparability of the R and r coefficients that should be noted. The (Pearson) r is used for relationships between two variables, while R is used in the context of multiple regression with one independent variable and multiple dependent variables [50]. One important difference is that R cannot be negative, while r can be [50]. Another is that R has no associated hypothesis testing procedure (e.g., there is no set of critical values). Despite these differences, since both are measures of linear correlations, their comparison should still be informative.

Once the three sets of correlations addressing 1) - 3) were obtained for each of the prototypes, these correlations themselves were tested for significant differences. Using the size of the Pearson and multiple correlation coefficients that were observed to be statistically significant, the Fisher's z values and their associated p-values were calculated, with the focus being on whether there are any statistically significant differences in these correlations. In the case of the coefficients of multiple correlation, they will be included in this comparison if the model it derives from produces at least one significant pairwise r. The prototype that showed the largest positive differences for largest number of statistically significant correlations (if such a difference is observed) was deemed the most accurate.

To answer the question regarding usability, the mean of the SUS scores for each prototype will be taken. A test for statistical significance among the three scores will then be performed. The prototype with the highest positive difference (if such a difference is observed) will be deemed the most successful in terms of usability.

The code implementing the data analyses described above can be temporarily accessed at the following URL for examination (it will be moved to a more suitable location once examination is past):

https://www.github.com/kseakgwa/Genet-Elicitation-Tool-Analysis-Scripts

4.5 Materials

The three high-fidelity prototypes, the LimeSurveys implementing the validation measures (which are in Appendix B), and the SUS scale (also implemented in LimeSurvey), which is also in Appendix B.

5. RESULTS

This chapter presents the descriptive statistics, the correlational analyses of the prototypes with the explicit statement and scenario validation tasks, as well as subsequent inter-validation task correlational analyses and internal consistency analyses.

5.1 Sample characteristics

In total, 40 participants interacted with the prototypes. Of these 40, 13 either did not complete the interaction with their assigned prototype, or did not pass the attention checks. These participants thus did not progress to the validation survey. Of the 27 who progressed to the validation survey, only 20 completed the survey. Thus, in total, 20 participants were included in the data analysis. 17 of these were from the population of students from the University of Cape Town in the computer science department enrolled in CSC1016 and CSC5014Z. The other subset (n= 3) of the sample was from the "Zaphil" group. Of these three respondents, two specialize in moral philosophy.

Of the 20 participants whose data was included in the analysis, three interacted with CogSci-PT, nine interacted with HCI-PT, and eight interacted with KE-PT. Of the three that interacted with CogSci-PT, all three were classified as utilitarian. Of the nine participants who interacted with HCI-PT, five were classified as utilitarian, and four as deontologists. Of the eight participants who interacted with KE-PT, four were classified as utilitarian, two as deontologists and two as egoists.

5.2 Measurement errors

There were some errors which occurred during the data collection phase, which were considered and compensated for when the analyses below were conducted. While these errors did not affect the number of participants whose responses were analysed, they did affect the number of analysable responses per participant.

The first kind of error that occurred was database entry error, which impacted the collection of the responses of participants who interacted with CogSci-PT and HCI-PT. These were due to a bug in the code of these prototypes that saved the same response value twice. The responses affected (i.e., one response per user of each of the affected prototypes) were removed from the dataset used in the analyses that follows.

There was also another bug in the code of HCI-PT that caused one of the questions to be only partially displayed to participants. Thus, responses to this question were also excluded from the dataset analyzed.

5.3 Descriptive statistics for the prototypes tested

This section reports the scores assigned for each moral theory for each user. It is segmented by prototype.

5.4 Descriptive statistics for CogSci-PT

For CogSci-PT the variables used for classification are the percentages scored on the items from OUS, the MFDA, and the ES. The descriptive statistics for these scores are summarized in Table 6 below.

Table 6 CogSci-PT OUS, the MFDA, and the ES descriptive statistics

User ID	OUS Percentage	MFDA Percentage	ES Percentage
12	76.19	8.89	54.00
15	63.49	8.89	34.00
17	58.73	37.78	57.00
Mean and SD	66.14 (SD= 9.03)	18.52 (SD= 16.68)	48.33 (SD= 12.50)

Here we can see that all CogSci-PT users' utilitarianism scores were their highest scores. Thus, all three CogSci-PT users were classified as utilitarians.

5.5 Descriptive statistics for HCI-PT

For HCI-PT the first set of variables used for classification are the percentages scored on the subscales from the MFQ. The other variable used is a dilemma score. The descriptive statistics for these scores are summarized in Table 7 below.

Table 7 HCI-PT descriptive statistics

The total possible scores for the variables are as follows: dilemma score total = 2.95, and the value score totals= 5

User ID	Care score	Fairness score	Loyalty score	Authority score	Sanctity score	Dilemma score
11	2.66	3.6	1.0	2.6	1.8	2.01
13	4.83	4.0	2.4	2.6	3.4	1.32
15	3.00	3.4	1.6	1.6	3.0	1.83
17	3.83	5.0	1.6	2.4	2.0	1.64
18	4.83	3.0	0.0	3.2	3.6	1.26
20	4.17	3.8	1.8	3.4	3.6	1.39
21	2.50	3.4	0.8	3.0	2.4	1.95
22	3.33	4.2	2.8	3.0	0.0	2.16
24	3.33	1.8	2.0	4.0	1.8	2.02
Mean and SD	3.61 (SD= 0.87)	3.58 (SD= 0.88)	1.85 (SD=1.09)	2.87 (SD= 0.68)	2.40 (SD= 1.17)	1.73 (SD= 0.34)

Based on these results, HCI-PT categorized five participants as utilitarians (specifically users 11, 15, 17, 21, 24) and the other four as deontologists (specifically users 13, 18, 20, 22).

5.6 Descriptive statistics for KE-PT

For KE-PT, the variables used for classification are the scores on the TEs presented by the prototypes. These scores are the deontology score, utilitarianism score, egoism score and DCT score. The descriptive statistics for these scores are summarized in Table 8 below.

Table 8 KE-PT descriptive statistics

The maximum score for all moral theory scores is 6.

User ID	Deontology score	Utilitarianism score	Egoism score	DCT score
64	4	1	0	1
65	3	2	1	0
66	1	1	4	0
70	0	6	0	0
71	0	6	0	0
72	0	6	0	0
74	0	5	1	0
75	0	2	4	0
Mean and SD	1 (SD=1.60)	3.63 (SD=2.33)	1.25 (SD=1.75)	0.13 (SD=0.35)

Based on these results, KE-PT categorized four participants as utilitarians (specifically users 70, 71, 72, 74), two as deontologists (specifically users 64 and 65) and two as egoists (specifically users 66 and 75).

5.7 Descriptive statistics for survey responses

This section reports the descriptive statistics for the validation measures.

5.7.1 *Results of moral philosopher view prediction*

As mentioned in Section 4.4, a cohort of ethicists was used to see if their self-reported preferred moral theory correlated with the moral theory assigned by the prototypes they interacted with. While three philosophers participated, only two specialized in moral philosophy.

One of these moral philosophers interacted with HCI-PT and the other with KE-PT. Both were classified as utilitarians. On the post prototype survey part of the study, when asked to self-describe their ethical outlook the ethicist who interacted with HCI-PT responded that they

were: “Probably something like a rule utilitarian who is quite permissive...”. The other ethicist said: “A kind of utilitarian virtue ethics/ethics of care”.

Thus, it seems that for both moral philosophers, the prototypes they interacted with gave them the nearest classification to their actual beliefs that was available to the prototype. This is evident because they were both categorized as utilitarian, while both their responses characterize them as kinds of utilitarian. So, HCI-PT and KE-PT were as accurate as possible given the moral theories available to them.

Unfortunately, due to this small sample size, generalizations about the accuracy of these prototypes cannot be made based on this data.

5.7.2 *Explicit statement response statistics*

In Table 9 below are the descriptive statistics of the responses to the explicit statement validation task, for CogSci-PT’s users.

Table 9 CogSci-PT explicit statement response statistics.

The responses to the explicit statements are on a 1-5 Likert scale.

User ID	Preferred moral theory	Explicit statement for utilitarianism	Explicit statement for deontology	Explicit statement for egoism	Explicit statement for DCT
12	Utilitarianism	4	2	4	2
15	Utilitarianism	4	1	1	1
17	Utilitarianism	5	4	3	4

Here we see that while all the participants had high utilitarianism scores based on their responses to the utilitarianism explicit statement (i.e., 4 -5 out of 5), users 12 and 17 also rated the explicit statement of at least one other moral theory with a score of at least 4. In one case (i.e., user id= 12), the other highly rated explicit statement was equal to the rating given to the utilitarianism statement.

In Table 10 below are the descriptive statistics of the responses to the explicit statement validation task, for HCI-PT’s users.

Table 10 HCI-PT explicit statement response statistics

The responses to the explicit statements are on a 1-5 Likert scale.

User ID	Preferred moral theory	Explicit statement for utilitarianism	Explicit statement for deontology	Explicit statement for egoism	Explicit statement for DCT
11	Utilitarianism	3	1	1	1
13	Deontology	1	2	1	1

15	Utilitarianism	2	3	2	3
17	Utilitarianism	3	4	1	1
18	Deontology	2	2	2	5
20	Deontology	1	4	1	2
21	Utilitarianism	3	1	3	1
22	Deontology	4	4	4	1
24	Utilitarianism	5	2	1	1

From the table above we see that, for the utilitarians, two respondents had the utilitarianism explicit statement score as their highest explicit statement score (user ID= 11 and 24). For these same participants, two had utilitarianism scores which were tied with another explicit statement score as their highest (i.e., user ID= 15 and user ID= 17). Another two had an explicit statement response other than utilitarianism as their highest score (user ID= 21 and user ID= 15)

For the deontologists, two had their deontological explicit statement score as their highest score (i.e., user ID= 13 and user ID=20), one had a deontological explicit statement response which was tied for their highest score (i.e., user ID=22), and one had a score other than deontology as their highest score (user ID=18).

In Table 11 below are the descriptive statistics of the response to the explicit statement validation task, for KE-PT's users.

Table 11 KE-PT explicit statement response statistics

The responses to the explicit statements are on a 1-5 Likert scale.

User ID	Preferred moral theory	Explicit statement for utilitarianism	Explicit statement for deontology	Explicit statement for egoism	Explicit statement for DCT
64	Deontology	4	5	2	1
65	Deontology	4	3	2	1
66	Egoism	3	1	4	1
70	Utilitarianism	3	3	2	3
71	Utilitarianism	4	2	1	1
72	Utilitarianism	4	1	1	1
74	Utilitarianism	4	2	4	2
75	Egoism	2	4	1	1

From the table above we see that, for the utilitarians, two had the utilitarianism explicit statement score as their highest explicit statement score (i.e., user ID= 71 and 72). The other two utilitarians had utilitarianism scores which were tied with another explicit statement score as their highest (i.e., user ID= 70 and 74).

For the deontologists, none had their deontological explicit statement score as their highest score (user ID = 64), none had a deontological explicit statement response which was tied for

their highest score, and one had a score other than deontology as their highest score (user ID = 65).

For the egoists, one had their egoism explicit statement score as their highest score (i.e., user ID=66), and one had a score other than egoism as their highest score (i.e., user ID=75).

5.7.3 *Validation scenario response statistics*

In Table 12 below are the descriptive statistics of the response to the scenario-based validation task, for participants who interacted with CogSci-PT.

Table 12 CogSci-PT Validation scenario response statistics

Scenario scores are reported as percentages.

User ID	Preferred moral theory	utilitarianism scenario score	deontology scenario score	egoism scenario score	DCT scenario score
12	Utilitarianism	77.78	10	75	20
15	Utilitarianism	55.56	50	37.50	20
17	Utilitarianism	55.56	50	12.50	90
Mean and SD		62.96 (SD=12.83)	36.67 (SD=23.09)	43.33 (SD=40.41)	41.67 (SD=31.46)

From the table we see that for the CogSci-PT participants, all of whom were classified as utilitarians by CogSci-PT, two of the three (i.e., user id= 12 and user id=15) of them had their utilitarianism scenario scores as the highest of scenario scores. One user (i.e., user ID= 17) had their DCT scenario score as their highest scenario score, with utilitarianism coming second.

In Table 13 below are the descriptive statistics of the response to the scenario-based validation task, for participants who interacted with HCI-PT.

Table 13 HCI-PT Validation scenario response statistics

Scenario scores are reported as percentages.

User ID	Preferred moral theory	utilitarianism scenario score	deontology scenario score	egoism scenario score	DCT scenario score
11	Utilitarianism	100	10	37.5	40
13	Deontology	55.56	40	12.5	50
15	Utilitarianism	55.56	50	25	90
17	Utilitarianism	100	10	12.5	20
18	Deontology	55.56	60	12.5	100
20	Deontology	33.33	60	75	100
21	Utilitarianism	77.78	10	62.50	20

22	Deontology	33.33	60	50	60
24	Utilitarianism	55.56	20	37.50	60
Mean and SD		62.94 (SD=24.84)	35.56 (SD=27.97)	36.11 (SD=22.92)	60 (SD=31.22)

From the table above we see that, for the utilitarians, three had the utilitarianism scenario score as their highest scenario score (i.e., user ID= 11, 17 and 21). Additionally, one had another scenario other than utilitarianism score as their highest (i.e., user ID= 24).

For the deontologists, none had their deontological scenario score as their highest score, one had a deontological scenario score which was tied for their highest score (i.e., user ID= 22)., and three had a score other than deontology as their highest score (user ID=13, 18 and 20).

In Table 14 below are the descriptive statistics of the response to the scenario-based validation task, for participants who interacted with KE-PT.

Table 14 KE-PT Validation scenario response statistics

Scenario scores are reported as percentages.

User ID	Preferred moral theory	utilitarianism scenario score	deontology scenario score	egoism scenario score	DCT scenario score
64	Deontology	50	11.11	50	50
65	Deontology	50	66.67	20	25
66	Egoism	10	44.44	60	75
70	Utilitarianism	40	44.44	30	62.5
71	Utilitarianism	30	44.44	10	50
72	Utilitarianism	50	44.44	40	25
74	Utilitarianism	20	44.44	70	25
75	Egoism	60	11.11	60	25
Mean and SD		38.75 (SD=17.27)	38.89 (SD=18.78)	42.5 (SD=18.71)	42.19 (SD=19.97)

From the table above we see that, for the utilitarians, one had the utilitarianism scenario score as their highest scenario score (i.e., user ID= 72). Three had a score other than the utilitarianism scenario score as their highest score (i.e., user ID=70, 71 and 74).

For the deontologists, one had deontology score tied with another for their highest (user ID = 64), while the other has a score other than deontology as their highest score (i.e., user ID= 65).

Finally, for the egoists, one had an egoism score tied with another for their highest (user ID = 75), while the other has a score other than egoism as their highest score (i.e., user ID= 66).

5.8 SUS Analysis

As mentioned in Section 4.4, the participants were also asked to complete the SUS scales following their completion of the two validation tasks.

The SUS score for each of the prototypes was computed based on the responses of the participants who interacted with it. These scores can be seen in Table 15 below:

Table 15 Prototype SUS scores

The SUS scores reported are mean scores for all the users of each prototype.

Prototype	SUS score	Usability status
CogSci-PT	79.2	Acceptable
HCI-PT	71.4	Acceptable
KE-PT	73.1	Acceptable

Here we see that all of the prototypes' SUS scores were above 68, and were thus high enough to meet the requirements of acceptability [49].

A set of two tailed two sample t-tests were run to ascertain if the differences between the SUS scores were statistically significant (at a confidence level of 0.95). Table 16 below presents the results of these t tests.

Table 16 SUS t-test comparison results

Prototypes compared	T	P
CogSci-PT and HCI-PT	0.63	0.54
CogSci-PT and KE-PT	0.68	0.51
HCI-PT and KE-PT	-0.23	0.82

Here we see that, for all the p-values, $p > 0.05$. Thus, there is no statistically significant difference between the SUS scores. Thus H2, cannot be rejected.

5.9 Validation measure correlations

This section reports the correlations between the prototype scores and the validation measures.

5.9.1 *Explicit statement correlations*

Table 17 below contains the correlations between the scores on the MFDA, ES and OUS scales used in CogSci-PT, and the corresponding responses given on the explicit statement validation task.

Table 17 CogSci-PT explicit statement correlation results

Variables being correlated	R	P
OUS score and Utilitarian explicit statement score	-0.71	0.50
OUS score and Deontology explicit statement score	-0.44	0.71
ES score and Deontology explicit statement score	0.90	0.29
MFDA score and DCT explicit statement score	0.94	0.21

As can be seen in the table above, there were no correlations that were statistically significant.

Interestingly, the utilitarian explicit statement scores and OUS score are moderately negatively correlated (r= -0.71), which was not observed in [11].

Regarding the users of HCI-PT, Table 18 below contains the correlations between the scores on the MFQ scale as implemented in HCI-PT, and the responses given on the explicit statement validation task. As explained in Section 4.2, Pearson's Rs (r) are reported for the utilitarian and deontological correlations, while for DCT and egoism the multiple correlation coefficient (R) is reported. In terms of the p-values for these, a series of t tests were used to produce the p-value for each of the dependent variables.

Table 18 HCI-PT explicit statement correlation results

Variables being correlated	r/R	Statistical significance
Dilemma score and Utilitarian explicit statement score	0.80	0.00
Dilemma score and Deontology explicit statement score	- 0.11	0.76
Loyalty score, care score, fairness score, authority score, sanctity score, dilemma score and Egoism explicit statement score	0.56	No significant t-test results
Loyalty score, authority score, sanctity score, dilemma score and DCT explicit statement score	0.75	No significant t-test results

Although all the results had the predicted directions, the utilitarianism correlation was the only statistically significant one (p= 0.00).

For the users of KE-PT, Table 19 below contains the correlations between the utilitarianism, egoism, DCT and deontology scores assigned by KE-PT, and the corresponding responses given to the explicit statement validation task.

Table 19 KE-PT explicit statement correlation results

Variables being correlated	R	p
Utilitarianism score and Utilitarian explicit statement score	0.28	0.49
Deontology and Deontology explicit statement score	0.57	0.14
Egoism score and Egoism explicit statement score	0.31	0.45
DCT score and DCT explicit statement score	- 0.30	0.62

As the table above shows, no statistically significant results were observed.

The DCT explicit statement responses negatively correlate with the corresponding scores assigned by KE-PT (r = -0.30). This was not predicted, since all the scores given by KE-PT should positively correlate with the validation measures.

5.9.2 *Scenario correlations*

Now for the the scenario validation measure results for the users of CogSci-PT. Table 20 below contains the correlations between the scores on the MFDA, ES and OUS scales used in CogSci-PT, and the scores on the scenario-based validation task.

Table 20 CogSci-PT scenario correlation results

Variables being correlated	R	p
Dilemma score and Utilitarian scenario score	0.96	0.17
Dilemma score and Deontology scenario score	-0.96	0.17
ES score and Egoism scenario score	-0.01	1.00
MFDA score and DCT scenario score	1.0	0.00

As the table above shows, only the correlation between the MFDA score and the DCT validation scenarios is significant (r= 1.0, p=0.00).

The correlation between the egoism score assigned by CogSci-PT and the scenario-based egoism score is negative (r= -0.01, p=1.00), despite the ES score and the egoism scenario score being predicted to positively correlate [19].

Now, t for the users of HCI-PT, Table 21 below contains the r and Rs describing the relationship between the scores on the MFQ scale as implemented in HCI-PT, and the scenario-based validation scores.

Table 21 HCI-PT scenario correlation results

Variables being correlated	r/ R	p
Dilemma score and Utilitarian scenario score	0.17	0.67
Dilemma score and Deontology scenario score	-0.39	0.29
Loyalty score, care score, fairness score, authority score, sanctity score, dilemma score and Egoism scenario score	0.86	No significant correlations(s)
Loyalty score, authority score, sanctity score, dilemma score and DCT scenario score	0.50	No significant correlations(s)

Here we see that no statistically significant results were observed. All the results showed the predicted direction**.**

For the users of KE-PT, Table 22 below reports the correlations between the scores on the utilitarianism, egoism, DCT and deontology scores assigned by KE-PT, and the corresponding explicit statement validation task scores.

Table 22 KE-PT scenario correlation results

Variables being correlated	R	P
Utilitarianism score and Utilitarian explicit statement score	-0.08	0.84
Deontology and Deontology explicit statement score	-0.11	0.80
Egoism score and Egoism explicit statement score	0.56	0.15
DCT score and DCT explicit statement score	0.15	0.71

As the above shows, there were no correlations that were statistically significant.

There was also a weak negative correlation between the utilitarianism scores and scenario

responses (r = -0.08).

5.10 Summary of correlational analyses

This section summarises the foregoing correlational analyses, focusing on the statistically significant results, and the correlations with unpredicted directionality.

5.10.1 *Significant results:*

For CogSci-PT, the only statistically significant result was the correlation between the DCT with the DCT scenarios (r= 1.00, p=0.00). For HCI-PT, the only significant correlation with the responses to the explicit statement of utilitarianism (r =0.80, p =0.00).

5.10.2 *Unpredicted correlations:*

The results for both the explicit statement and scenario-based validation task as correlated with the scores assigned by the prototypes which display unpredicted directionality are presented in Table 23 below.

Table 23 Unpredicted negative results

Prototype	Moral theory	Validation type	R	p
CogSci-PT	Util	Explicit statement	-0.71	0.50
CogSci-PT	Ego	Scenario	-0.01	1.00
KE-PT	DCT	Explicit statement	-0.30	0.62
KE-PT	Util	Scenario	-0.08	0.84
KE-PT	Deo	Scenario	-0.11	0.80

This data will be discussed in more detail in Section 5.12 below.

5.11 Comparing correlation size

To compare the size of the Pearson coefficients that were observed to be statistically significant in the foregoing analyses, the Fisher's z values and their associated p-values were computed.

The statistically significant correlations produced by the CogSci-PT and HCI-PT participants were the only ones compared. The result of this comparison was z= 1.02 and p=0.31.

It should be kept in mind that the sample size of both CogSci-PT and HCI-PT users is smaller than is usually recommended for the Fisher z transform [51]. So, the results reported in the previous paragraph may not reflect the true difference in efficacy of the prototypes being compared.

Another issue is that the correlations being compared are based on two sets of completely different variables. Specifically, one measures the relationship between CogSci-PT scores and the scenario validation scores, while the other measures that between HCI-PT and explicit statement scores. Thus, while possibly informative, how to interpret the results of the transform

is less clear than for correlations with scores of the same kind.

Keeping the caveats above in mind, it can tentatively be said that between the only two statistically significant correlations, there was no statistically significant difference.

5.12 Discussion of summary tables

So, two prototypes detect one strong ($r>0.79$) and statistically significant correlation. For HCI-PT, the correlation is with the explicit statement validation scores. For CogSci-PT, the significant correlation was with the DCT validation scenarios

Thus, only these statistically significant correlations of CogSci-PT and HCI-PT could be compared using the Fischer's Z method, and this comparison yielded no significant results.

Regarding the correlations with unpredicted directionality, HCI-PT is the only prototype which did not produce any such results, even when one takes into account that the DCT and egoism Rs could not produce negative results for reasons discussed in section 4.2. CogSci-PT utilitarianism scores strongly negatively correlate with ratings of agreement with its corresponding explicit statement. CogSci-PT egoism scores also weakly negatively correlate with judgements in scenarios where acting egoistically is an option. KE-PT has similar results, with the correlation of DCT scores with its corresponding explicit statement. The KE-PT correlations of the deontology and utilitarianism scores with the corresponding scenario scores also showed unpredicted direction.

From this data, one cannot reject the null hypothesis for H1, which is the hypothesis that there isa significant difference between the performance of the prototypes on the validation tasks.

Yet, if we look at these results in terms of which prototype had the best ratio of statistically significant to unpredicted correlations, HCI-PT performs the best, with one statistically significant result and no correlations with unexpected directionality. On the other hand, KE-PT performs the worst, with no statistically significant correlations and three correlations with unpredicted directionality.

The subsequent sections in this chapter will detail a series of additional rounds of analysis, each meant to provide further data for explaining (in the discussion chapter) the performance of all of the prototypes.

5.13 OUS-IB and OUS-IH analysis

One approach for checking the robustness of the observed correlational results, at least in CogSci-PT, is to check if they vary if one treats the OUS as a composite of two subscales, despite the critiques in [21]. [1] As explained in Section 2.3, these subscales measure what Kahane et al call "impartial beneficence" (using the OUS-IB subscale) and "instrumentalharm" (using the OUS-IH subscale) [11].

In the sections below, the CogSci-PT OUS subscale scores and validation scenario score correlations are reported. These correlations are between the scores of the participants who interacted with CogSci-PT, when their total scores are decomposed into OUS-IB and OUS-IH scores.

[1] While authors of MFDA say there are 3 conceptual concerns tracked by the scale, they treat them as all tracking the same single construct [14] and [18]. The authors of the ES also take it to be measuring one construct [19].

5.13.1 *OUS-IB and the Validation Scenarios*

The correlations between the OUS-IB scores, the deontology and utilitarianism the validation scenarios scores will now be presented. The correlation of OUS-IB and the deontology scenario has a magnitude of r = -0.76, with a p-value of p= 0.45. While the utilitarianism correlation is r=0.76 with a p-value of 0.45.

Thus, there is a lower r for the correlation between OUS-IB and the utilitarianism validation scenario score, than between the total OUS scale and the utilitarian validation scenarios as reported in Table 20 (from r=0.96 to r=0.76). The p-value is also lower (from p=0.50 to p=0.45).

For the deontology validation scenario score, we see another decrease in r when compared to the correlation of those same scenarios with the total OUS score (from r= -0.96 to r=-0.75) in Table 20. There is also an increase in p-value (from p=0.10 to p=0.45).

There is also a positive correlation between the utilitarian scenarios and the OUS-IB scores, and a negative correlation with those same scores and the deontological scenarios. Thus, there are no correlations with unpredicted directionality. This the same as the directionalities reported in Table 20.

5.13.2 *OUS-IH and the Validation Scenarios*

We now turn to the correlations of the OUS-IH scores with the deontology and utilitarianism validation scenario scores. The correlation of the deontology scenarios with the OUS-IB scores is r = - 0.96, with a p-value of p= 0.10. While the correlation with the utilitarianism scenarios is r= 0.99 with a p-value of 0.45.

Thus, r does not change when correlating either the OUS-IH or the total OUS scale (reported in Table 20) with the utilitarian validation scenarios (from r=0.96 to r=0.96). The p-value is also the same (from p=0.10 to p=0.10).

For the deontology validation scenario score correlation, we see another constant magnitude when compared to the correlation of those scenarios with the total OUS score as reported in Table 20 (i.e., r= -0.96 to r=-0.99). The p-value has also stayed the same (from r=0.10 to r=0.10).

Finally, we can also see there is a positive correlation between the utilitarian scenarios and the OUS-IH scores, and a negative correlation with those same scores and the deontological scenarios. Thus, there are no correlations with unexpected directionality. This is consistent with the directionality of the correlations of the scenarios with the total OUS scores as reported in Table 20.

5.13.3 *OUS-IB and the Explicit Statements*

This section concerns the correlation between the OUS-IB scores for CogSci-PT users and the deontology and utilitarianism explicit statement scores. The correlation for deontology explicit statement responses with the OUS-IB has a magnitude of r = -0.94, with a p-value of p= 0.21. While the utilitarianism correlation is r = -0.79 with a p-value of 0.42.

As the results above show, there is an increase in the r for the correlation between OUS-IB and the explicit utilitarian statement validation compared to the total OUS scale and this statement

(from r=-0.71 to r=-0.79) as reported in Table 17. The p-value also decreased (from p=0.50 to p=0.42).

For the deontology score correlation, there is another increase in r (from r= -0.44 to r=-0.94) when compared to the results in Table 17. The p-value has decreased (from r=0.71 to r=0.21).

From the preceding paragraph we see there is a negative correlation between the utilitarianism explicit statement and the OUS-IB scores, and a negative correlation with those same scores and the deontological explicit statement. This is consistent with the directionality of the explicit statement with the total OUS scores in Table 17.

5.13.4 *OUS-IH and the Explicit Statements*

Now the correlations between OUS-IB scores, the deontology and utilitarianism explicit statement scores will be discussed. The correlation of the deontology explicit statement score and the OUS-IH has a magnitude of r = -0.36, with a p-value of p= 0.77. While the utilitarianism explicit statement correlation is r = -0.03 with a p-value of 0.98.

From the above we can see that there is a decrease in the r for the correlation between OUS-IH scores and the explicit utilitarian statement, when compared to the correlation between the total OUS scale, and the explicit utilitarian statement as reported in Table 17 (from r=-0.71 to r=-0.03). The p-value increased (from p=0.50 to p=0.98).

For the correlation between the deontological explicit statement scores and the OUS-IH scores, there is another decrease in the magnitude of r, when compared to the correlation between the total OUS scale and the explicit deontology statement score as reported in Table 17 (from r= -0.44 to r=-0.36). The p-value has increased (from r=0.71 to r=0.77).

Finally, we can also see there is a negative correlation between the utilitarianism explicit statement and the OUS-IB scores, and a negative correlation with those same scores and the deontological explicit statement. This is consistent with the directionality of the explicit statement with the total OUS scores in Table 17.

5.13.5 *Summary of the subscale-based analysis*

This subscale-based analysis reveals that there were mostly decreases in the magnitude of the correlations of the OUS-IB and OUS-IH scales with the validation measures when compared to the initial (total) OUS-based analysis. Of the eight subscale correlations, four yielded decreases in the size of the correlation, when compared to the correlations of those same validation scores with the total OUS scale in Tables 16 and 19. Two of these came from the OUS-IH-explicit statement analysis, and the other two come from the OUS-IB-scenario analysis. The correlations between the OUS-IH scores with the deontology and utilitarianism scenario scores both yield no change in r when compared to the analysis in Table 20. Finally, the OUS-IB-explicit statement correlations yielded two increases, when compared to the initial analysis in Table 17.

The statistically significant result observed in the initial analysis in Table 20 was also not observed in the subscale-based analyses. The unpredicted negative correlations between the total OUS explicit statement scores and the utilitarianism explicit statement scores (see Table 17) also persist on both the OUS-IB and OUS-IH analyses.

Given that the only increases in r were in the correlations with unpredicted directionality, this suggests the results observed in the initial analyses in Tables 17 and Table 20, in terms of the

unpredicted directionality and size of the correlations, were not due to a failure to follow [21] in subdividing their scale. The initial significant result was also not replicated.

One possible explanation for these findings is that they are artefacts caused by the small sample size. The next section will investigate this possibility.

5.14 Power analyses

The sample size of users of each prototype ranges from three to nine, which is much lower than the sample sizes of the studies which yielded the scales/elicitation stimuli used (see [11], [14], [19]). So, one possible explanation for the findings of the initial analyses is this difference in sample size. This is because with a small sample size one needs a much higher r for a particular correlation to have a p-value low enough to reject the null hypothesis (whether it be at alpha=0.05 or any other confidence level) [52]. This means a higher chance of not rejecting the null hypothesis even when there is an effect in the population (i.e., type II error).

To see whether the sample sizes in this study were too small to mitigate this problem, a set of post-hoc power analyses were employed.

Thus, power analyses were used to calculate sample sizes with sufficient power for accurately detecting correlations of r= 0.7 to r=0.9. This was done while assuming a statistical power of 0.8 and an alpha of 0.05 (which are both the standard in the literature) [52]. This yielded the results in Table 24 below.

Table 24 Power analyses results

R	0.7	0.8	0.9
Sample size	13.43	9.50	6.62

For CogSci-PT, n= 3. For all the correlation sizes measured, the samples sizes that would give a power of 0.8 are >3. Thus, the power of Pearson's R procedure based on the sample of participants that interacted with CogSci-PT is below the traditional acceptability standard (i.e., 0.8) for all correlation sizes investigated.

The sample size for HCI-PT (n= 9) on the other hand, is sufficiently large to be likely to correctly reject the null hypothesis, if the correlation size in the population is 0.9 or above. So, since this would yield a power of 0.8, there is only a 20% chance that the insignificant results observed are failing to detect correlations of 0.9 or more.

Finally, KE-PT (n=8) is also sufficiently large to be likely to correctly reject the null hypothesis, if the correlation size in the population is 0.9 or above. So, there is only a 20% chance that the insignificant results observed are not detecting correlations of 0.9 or more as well.

So, the low number of significant correlations could be caused by this small sample size which is unable to reliably detect smaller correlations (i.e., r<0.9), at least in the case of HCI-PT and KE-PT. In the case of CogSci-PT though, the probability of false negatives is even higher, given that the sample size of participants who interacted with it is small enough for the correlational analysis to output false negatives even for correlations in the population which are up to 0.9.

5.15 Inter-validation measure correlation.

One way to test whether the sample size is in fact not allowing the detection of correlations which are significant but which are small, is to check how the responses on the two validation tasks correlate with each other. This is because this inter-validation task correlation analysis allows for a bigger sample (n=20)[2] than just each subset based on the prototype they interacted with.

This inter-validation task analysis will contribute to showing whether observed results are due to features of the prototype, the measurement errors, or the features of the participants in the sample (e.g., participants not having one moral theory).

Below, in Table 25, are the results of the Pearson R correlations between ratings of agreement with the explicit statements of the moral theories studied, and the moral scenarios used for validation.

Table 25 Inter-validation measure correlation

Moral theory	R	P
Deontology	0.04	0.96
Egoism	0.44	0.05
DCT	0.67	0.00
Utilitarianism	-0.04	0.87

Here the results are a mix of low to moderate correlations. The one which is statistically significant is the correlation between the DCT scenarios and the DCT explicit statement (p= 0.00). The unpredicted negative utilitarianism correlation (r = -0.04) also persists from the results reported in Table 23 for the CogSci-PT utilitarianism explicit statement (r = -0.71) and KE-PT utilitarianism scenario (r = -0.08) correlations, but there is no unpredicted directionality for egoists, as there was for the correlation of the CogSci-PT egoism score with the egoism scenario score as reported in Table 23. Lastly, there was a general reduction of the size of the Pearson's Rs in the inter-validation analysis when compared to the correlation sizes of the prototype scores with the validation measures. This effect was largest in the utilitarianism and deontology correlations, since none of the tables in section 5.9 have utilitarianism or deontology correlations lower than in Table 25 above (r=0.04 and r= -0.04 respectively). 50% of the egoism correlations were also reduced when compared to the data reported in that same section. The DCT correlations were the only set of results where most correlations were not reduced, with 33% of the original correlations being lower than that in Table 25.

5.16 Cronbach's alpha

In all the correlational analyses in Section 5.9, we observed correlations that are mostly statistically insignificant, and the persistence of unpredicted directionality of some of these correlations. Both issues could be the result of inconsistency of the participants. In this context, an inconsistent user would be a user that does not have one moral theory that guides them in their moral judgements or at least responds as if they do not have one. One way of assessing

[2] Big enough to detect an r of 0.60 or above at alpha=0.05, with statistical power of 0.8.

whether there is one such theory underlying each participant's judgements (at least as collected by responses to items on a psychometric scale), is by using a measure of reliability/internal consistency.

In social scientific research, reliability is a property of psychometric instruments whereby the instrument elicits consistent responses from the participant [53]. The kind of reliability most relevant for our purposes is internal consistency, which is the consistency of the responses of a participant over multiple items on a scale. By measuring internal consistency, we can tell whether the pattern of responses of a participant to the questions encountered while interacting with a prototype is congruent with the hypothesis that one psychological construct (e.g., a moral theory) underlies those responses. We can also do the same for the scenarios presented to participants as a validation measure. This would only be an indirect test, since knowing that one psychological construct underlies a set of responses does not allow you to make inferences about what the nature of this construct is. But to have one moral theory underlying participants' responses, one must have one psychological construct underlying those responses. So, if it is observed that one psychological construct does not underlie the responses, we can infer that one moral theory does not underlie those responses nor the responses on the validation scenarios.

A popular measure of internal consistency [53] is Cronbach's alpha (CA). CA measures the extent to which a set of questions are tracking one underlying construct. An acceptable CA is 0.7 or above [53].

The results of the CA analyses are reported below[3], starting with the CAs for the validation scenarios.

5.16.1 *CA for the validation scenarios*

The CA for the DCT validation scenario responses is 0.63 with a confidence interval of 0.06 – 0.085. The Cronbach's alpha for the egoism responses is 0.45 with a confidence interval of -0.38 – 0.78. Thus, both the Cronbach's alphas for DCT and egoism are below 0.70, and are therefore below the threshold for acceptable reliability.

Cronbach alphas for the deontology and utilitarian scenarios could not be computed because there is variance across participants in the numbers of utilitarian and deontological answers, and the Cronbach's alpha computation requires an equal number of answers across participants.

5.16.2 *CAs for the prototypes*

The tables in the sections below report the CAs for each of the prototypes.

5.16.2.1 CogSci-PT CA

In Table 26 below are the CAs for the responses to CogSci-PT. These are segmented by the scales/subscales the questions used in the CogSci-PT are drawn from. [4]

[3] No intra-explicit-statement Cronbach's alpha could be calculated because there is only one explicit statement per moral theory.

[4] Note that one response to the OUS overall, one response to the OUS-IH, and two responses to the ES were excluded due to the data entry error causing these items to have only two responses each.

Table 26 CAs for CogSci-PT

Scale	CA	Confidence interval
OUS Total	-1.42	-10.76, 0.94
OUS-IB	-1.70	-15.34, 0.93
OUS-IH	0.85	-0.64, 0.10
MFDA	0.71	-0.76, 0.10
ES	0.84	0.35, 0.10

Here the CA of the OUS total (i.e., including both OUS-IH and OUS-IB items) and OUS-IB taken alone are both below the acceptance threshold. OUS-IH, MFDA and ES all had acceptable CAs.

5.16.2.2 HCI-PT CA

In Table 27 below are the CAs for the responses to HCI-PT, segmented by the subscale of the MFS (i.e., value) the questions are drawn from. Due to the data entry error, there were two groups of answer sets for some of the subscales. These groups of sets are 1) groups of answer sets that have nine responses, and 2) groups of answer sets that have eight responses. These two sets were analysed separately since the CA computation requires the comparison of sets of answers that have the same number of answers.

Table 27 Cronbach's Alphas for HCI-PT

Value	CA (nine response sets)	Confidence interval (nine response sets)	CA (eight response sets)	Confidence interval (eight response sets)
Authority	0.21	-1.19, 0.80	NA	NA
Care	-1.31	-6.80, 0.50	0.67	-0.04, 0.92
Fairness	0.81	0.46, 0.95	NA	NA
Loyalty	0.81	0.06, 0.96	0.23	-2.41, 0.83
Sanctity	0.68	0.11, 0.92	NA	NA

Here the CA of the Care, Authority, and Sanctity subscales for the sets of responses that have nine answers are all below the acceptance threshold. Fairness, and Loyalty all had CAs that met the standard of acceptable reliability for response sets with nine answers. No eight response sets were acceptable.

5.16.2.3 KE-PT Cronbach's alpha

In Table 28 below are the CAs for the responses to KE-PT, segmented by the moral theory the TEs responded to were testing for.

Table 28 Cronbach's Alphas for KE-PT

Moral theory	CA	Confidence interval
Utilitarianism	0.55	-1.24, 0.91
Deontology	-0.55	-1.24, 0.91
DCT	0.67	-0.67, 0.93
Egoism	0.43	-1.84, 0.89

From the table we can see that none of the response sets hade CAs that were acceptable

5.16.3 *Summary of the CA analyses*

The acceptable scores are presented in the table in the sections below.

5.16.3.1 CogSci-PT

In Table 29 below are the acceptable Cronbach alphas for the responses to CogSci-PT, segmented by the scales the questions used in the CogSci-PT are drawn from.

Table 29 Acceptable Cronbach's for CogSci-PT

Scale/subscale	CA	Confidence interval
OUS-IH	0.85	-0.64, 0.97
MFDA	0.71	-0.76, 0.99
ES	0.84	0.35, 1.00

While all the Cronbach's alphas are acceptable, they are all also accompanied by wide confidence intervals, which indicate a high probability that the CA would not be observed again if retaken with another sample.

5.16.3.2 HCI-PT

The acceptable CAs for the responses to HCI-PT, segmented by the subscale of the MFS (i.e., value) the questions are drawn from will now be discussed. For the fairness value, the CA was 0.81 with a confidence interval of 0.47 – 0.95, will the CA for the loyalty subscale was also 0.81 with a confidence interval of 0.06 – 0.96.

While all of the CAs are acceptable, they are all also accompanied by wide confidence intervals, which indicate a high probability that the CA would not be observed again if retaken with another sample.

5.17 Analysis by category

The foregoing correlational analyses each take the scores of all participants who interacted with a prototype. They measure the correlation of the prototype score for each theory with their explicit statement and scenario counterparts, irrespective of the theory each participant was categorized as preferring.

A different approach is to sort the participants according to which prototype they interacted with, and the moral theory they were classified as preferring. One can then calculate the correlations of those who were classified as preferring a particular theory with the responses given to the prototypes and to items on the validation measures meant to measure that theory only.

This may result in higher correlations and with a lower p-value, as only the highest score each participant was given will be included in the correlation computation with only the validation measure items that are theorized to correlate with it.

The tables in the sections below report the Pearson correlation coefficients between utilitarianism score or the deontology score of the HCI-PT users (n=9) and the corresponding scenarios and explicit statement scores, depending on whether they were classified as utilitarians (n=5) or deontologists (n=4). The same tables also report the utilitarianism score, the deontology score, or egoism score of the KE-PT users (n=8), depending on whether they were classified as utilitarians (n=4), deontologists (n=2), or egoists (n=2).

5.17.1 *Validation Explicit Statement correlations*

Table 30 below reports the correlations of the HCI-PT and KE-PT scores of users classified by moral theory with the corresponding validation explicit statement responses.

Table 30 Explicit statement correlations by category

User classification	HCI-PT		KE-PT	
	R	p	r	p
Utilitarian	0.45	0.44	-0.33	0.67
Deontologist	0.66	0.34	1.0	1.0
Egoist	-	-	NA	NA

Note that all comparisons in the following discussion are to Table 18 for HCI-PT results, and to Table 19 for KE-PT results.

As the table above shows, the correlation of the utilitarianism scores of HCI-PT users who were classified as utilitarian with their utilitarian explicit statement scores is (r=0.45). For the KE-PT users this same correlation is (r=-0.33). The HCI-PT utilitarian explicit statement ratings correlation is than theanalogous correlation in the original analyses (r=0.80), while for the KE-PT users this is stronger than the original correlation (r=0.28). For the utilitarian users of HCI-PT, the p-value has also increased (p=0.44 from p=0.00), and is now above the significance threshold. For the KE-PT users, the p-value has also increased (p=0.67 from p=0.49). The same analysis for the deontology scores of HCI-PT deontologists shows that the correlation strength increased (r=0.66 from r=-0.11) with a decrease in p-value as well (p=0.34 from p=0.76). While for KE-PT users, the deontology scores of deontologists show that the correlation strength increased.

(r=1.00 from r=0.57), with an increase in p-value as well (p=1.00 from p=0.14). For KE-PT users classified as egoists, a correlation could not be computed due to the two ego scores assigned by KE-PT to the egoists being the same.

From the table above we can also see there is a positive correlation between the HCI-PT users' utilitarian explicit statement responses and their utilitarianism scores, as well as between the deontologists' scores and explicit statement responses. Thus, the HCI-PT deontology correlation displays unpredicted directionality. This is different from the directionality of the correlation correlations in the original analyses. For KE-PT users, the deontology correlation also has unpredicted directionality, but this is the same as the original correlation.

5.17.2 *Validation Scenario correlations*

Table 30 below reports the correlations between utilitarianism score, the deontology score or egoism score, and the validation scenario scores of the users depending on their classification.

Table 31 Scenario correlations by category

User classification	HCI-PT		KE-PT	
	R	P	r	P
Utilitarian	0.31	0.66	0.77	0.23
Deontologist	0.66	0.34	-1.0	1.0
Egoist	-	-	NA	NA

Note that, unless otherwise stated, all comparisons in the following discussion are to Table 21 for HCI-PT results, and to Table 22 for KE-PT results.

As the table above shows, the correlation of the utilitarian scores of HCI-PT users who were classified as utilitarian with their scenario scores is (r=0.31). For the KE-PT utilitarians this correlation is (r=0.77). For the HCI-PT utilitarians present score is more than the correlation between utilitarian scores of all HCI-PT participants with the same scenario responses (r=0.17). While for the KE-PT users it is stronger than the original correlation (r=-0.08). For the utilitarians of HCI-PT, the p-value has decreased (p=0.66 from p=0.67) and is still above the significance threshold. For the KE-PT utilitarians, the p-value has also decreased (p=0.23 from p=0.84). The same analysis for the deontology scores of HCI-PT deontologists shows that the correlation strength increased (r=0.66 from r=-0.39), with an increase in p-value as well (p=0.34 from p=0.29). While for KE-PT users, the deontology scores of deontologists show that the correlation strength increased (r=-1.00 from r=-0.11), with an increase in p-value as well (p=1.00 from p=0.80). For KE-PT egoists, a correlation could not be computed due to the two egoism scores assigned by KE-PT to the egoists being the same.

From the table we can also see there is a positive correlation between the HCI-PT users scores and the explicit statement responses, and the same deontological for the correlations. Thus, the HCI-PT deontology correlation displays unpredicted directionality. This is different from the directionality of the original correlations. For the KE-PT user results, while the utilitarianism correlation direction is different to the original correlation, the negative deontology correlation is the same the original.

Interestingly, for both HCI-PT and KE-PT deontologists, the scenario (in Table 31) and explicit statement (in Table 30) correlations are identical, with only the direction of the scenario correlations for those KE-PT deontologists differing from their explicit statement correlation ($r = -1.00$ and $p = 1.00$, from $r = -1.00$ and $p = 1.00$). This suggests that there is an underlying psychological construct being measured by both prototypes, which remains consistent in magnitude across validation tasks. This in turn suggests that despite the lack of significance of the relevant correlations, they do not represent an illusory relationship.

6. DISCUSSION

This chapter will first discuss the hypotheses given the results observed in the previous chapter. It then discusses the results of the validation studies of the scales used in the first prototype, particularly [11], in comparison to what was observed in the present study. The chapter then presents an interpretation of the results of the various analyses reported in Chapter 5. A summary of what the results from Chapter 5 show about the prototypes that were tested will then be given. Implications of these findings for the general project of eliciting users' moral theories will then be discussed. Finally, the limitations of the study will then be addressed.

6.1 The status of the hypotheses in light of the results

H1 (see Section 1.2) states that there is a significant difference between the computational instantiations of psychometric scale-based approaches which are explicitly formulated in analytic moral philosophical terms (representing cognitive science), psychometric scale-based approaches which take values as their unit of analysis when delivered via VSs (representing HCI), and proximity scaling approaches delivered via thought experiments (representing knowledge engineering) with regard to how accurately they are able to categorize users according to their moral theory. By this standard, given the results of the correlational analyses, there is no clear best performing prototype, as the only comparable results yielded no statistically significant difference when compared. Given this, the results of the present study provide insufficient evidence to reject the null hypothesis.

While there is no statistically significant difference between the performance of the prototypes, given that HCI-PT is the only one which did not yield any unpredicted correlations, it in this sense the best performing prototype. By this standard, KE-PT performed the worst, as it produced no statistically significant results with the validation measures, and produced three correlations with unpredicted directionality.

Given the failure to reject the null hypothesis for H1, while there are two significant results, the number of these results fall well below what would be needed for any of the prototypes to fulfill the requirements, which are to accurately classify users into (at least) all of the moral philosophical categories explicitly considered in [1].

H2 states that there is a significant difference between the computational instantiations of, psychometric scale-based approaches which are explicitly formulated in analytic moral philosophical terms (representing cognitive science), psychometric scale-based approaches which take values as their unit of analysis when delivered via VSs (representing HCI) and proximity scaling approaches delivered via thought experiment (representing knowledge engineering) with regard to usability. The analysis of the SUS results showed that, while all the prototypes had an acceptable SUS score, comparing these scores yielded no statistically significant differences. Thus, the null hypothesis that accompanies H2, cannot be rejected. So, it seems that all the prototypes have roughly equivalently acceptable user interfaces (at least as measure by the SUS). This also suggests that the small number of statistically significant correlations observed during validation are not due to usability issues.

6.2 What the scale developers observed

One important thing to consider when interpreting the results reported in the Chapter 5 is what the developers of psychometric scales used in CogSci-PT and HCI-PT observed, and how they

conceptualize lay moral cognition, since they had access to much larger sample sizes than the present study (i.e., n = 960).

As noted in Section 2.3, [11] contend that laymen do not have a single preferred moral theory which drives judgement and behaviour. They take the low correlations they observed between the scores on their scale and their validation measures (see Section 3.3 for more detail on their findings) to support their view.

It should be noted the findings in [11] differ from those observed in the present study in that in the present study the correlations with the validation measures were often moderate to large, but most were statistically insignificant ($p<0.05$). However, the observation of some negative correlations in both the various correlational analyses of the prototype assigned scores and the validation measures, and the inter-validation correlational analyses, also indicate a kind of contradictoriness in moral judgement which is also compatible with the hypothesis that lay moral psychology does not depend on a single moral theory.

6.3 Explaining the OUS subscale correlation analysis outcomes

In Section 6.15 which looks at the OUS-IB and OUS-IH results from CogSci-PT and their correlations with the validation measure scores, there was no improvement in the number of significant correlations (0) when compared to the original OUS-validation measure correlations (also 0) as reported in Section 5.9. While most correlations here had the predicted directionality, as also observed in [11], the negative correlation between the OUS scores (on both subscales) and the utilitarianism explicit statement rating persisted. This was not observed in [11], despite the utilitarianism statement being the same. This suggests inconsistency in the responses given by the participants above what was observed in [11]. Section 6.4, discusses a possible explanation for these and other findings.

6.4 Sample size as a factor explaining the results

One possible explanation for the lack of statistically significant correlations between the prototype-assigned scores and the validation measures scores, and the unpredicted directionality of some of these correlations, is the small sample size of participants that interacted with each prototype (i.e., n = 3-9). As the power analysis showed in Section 5.14, the lowest correlation that could be accurately detected with the sample size of any of the prototypes was r=0.9. Given that this is a correlation much stronger than the correlations observed in [11], the low number of significant correlations could be caused by these small sample sizes, which are not able to accurately detect smaller correlations, at least in the case of HCI-PT and KE-PT. In the case of CogSci-PT though, the probability of false negatives is higher than in the other prototypes, since n=3 is too small to accurately detect even correlations of r=0.9. This suggests that the statistically insignificant results for CogSci-PT are likely due to this lack of statistical power.

While the analysis above deals mostly with type II errors (i.e., false negatives), the small samples sizes generally, but particularly in the case of CogSci-PT, increases the probability of issues that could lead to type I errors (i.e., false positives). The most relevant of these issues for the data in this study is that the small sample size, especially when combined with a non-random sampling procedure, results in limited generalizability and thus, the analysis of data yielded from them is more informative about the participants in the sample, than the population they are drawn from. This could explain why unexpected directionality of some of the

correlations were observed in this study, but not by the scale developers who used larger samples.

6.5 Explaining the inter- validation correlation results

The persistence of mainly statistically insignificant correlations, even when considering the increased sample size afforded by the inter-validation dataset (i.e., n=20), is consistent with the sample size still being too small to detect statistically significant, but small, correlations. Evidence suggestive of this can be seen in the smaller correlations that are observed in this analysis. These findings are consistent with the idea that the actual size of the effects in the population are smaller than the prototype data suggests, and more in line with what the data collected by [11] demonstrates.

Interestingly, the intra-validation correlational analysis also yields a statistically significant DCT correlation, which is moderately strong. This is in line with the findings for both CogSci-PT and KE-PT scenario and explicit statement correlations respectively. Since none of the participants were classified by any of the prototypes as subscribing to DCT, this indicates that there were predominately low DCT scores assigned by the prototypes. On the validation measures, there was a 30.33 (out of 100) mean for the scenario-based DCT scores across all participants and a 1.7 mean score (out of 10) for agreement with the explicit statement of the definition of DCT. This pattern of low DCT scores across the prototypes and validation measures, along with the moderate to high correlation between the prototype DCT scores and the validation measures, as well as the inter-validation DCT correlations, supports the conclusion that there were strong anti-DCT tendencies in the sample. Since the sample was drawn from computer science students and philosophers, this strong anti-DCT signal could be due to a high prevalence of atheism in these populations generally [54].

Finally, note that the lack of significant results and the unexpected directionality in the prototype and validation score correlational analyses are not caused by the measurement errors that affected the data collection, since similar results are observed between the two validation measures, which were unaffected by the errors.

6.6 Explaining the CA results

The sections below discuss possible explanations for the CAs observed in Chapter 5.

6.6.1 *Validation CA*

For the internal consistency of the validation scenarios, both the CAs for DCT and egoism are below 0.70 in Section 5.16.1 and are therefore below the threshold for acceptable reliability. This suggests either possible issues with scenario selection/construction, since the collection of scenarios used were not pre-validated, or actual inconsistency in the responses given by participants. Given the other evidence presented in Chapter 5, such as the insignificant and unpredicted direction of (some of) the correlations of the inter-validation measures, the latter is the best supported explanation.

6.6.2 *Prototype CA*

Regarding the consistency of the responses to the stimuli given by the prototypes, CogSci-PT one shows the most reliability, despite the total OUS and OUS-IH having low CAs. This is in line with the findings of the respective scale developers (e.g., [11],[14]). HCI-PT's content is modified from validated prompts, which could explain the few acceptable CAs observed. This is because the assurances that accompany the content of CogSci-PT are not present in HCI-PT[5] and KE-PT[6], since the elicitation stimuli used in these were not empirically validated before their use. So, if the psychometric studies are taken as more trustworthy than the present results, because of their larger sample size, then one of the inferences drawn is that, since the inconsistency is not in the scales, it is in the participants or is an artifact caused by the small sample size, or there is something wrong with the validation measures, or all of these in some combination.

6.7 Category analysis

The category-based analysis shows no significant results. The initial unpredicted directionality of some of the correlations are also not changed. This could be due to further sub-division of the total size of the sample. This creates even smaller sample sizes than in the initial correlational analyses, and thus requires an even higher correlation to reach statistical significance.

6.8 Implications for approaches tested

Each prototype was constructed as a representation of an elicitation approach from a particular discipline. CogSci-PT represented the moral psychological approach, HCI-PT represented the HCI/VSD approach, and KE-PT represented the expert systems approach. Thus, the performance of the prototypes is also suggestive of the efficacy of these approaches.

Given the failure to reject the null hypotheses for H1 and H2, the investigation failed to show that any of these approaches are better or worse than the others, at least by the initial criteria set out in Section 4.4. However, as noted in Section 6.1, if the number of statistically significant results compared to results with unpredicted directionality is the criterion, then the VSD approach works best, while the expert systems approach works the worst.

This is interesting, since HCI-PT involved a novel combination of the MFQ and VSs, which unlike the scales used in CogSci-PT, had not been independently validated (although the MFQ had). The failure of KE-PT is less surprising, since none of the content used had been empirically validated. However, what this implies for the multi-dimensional and proximity scaling approaches used in expert systems and adapted for moral elicitation in [39], is of minimal impact. This is because, as discussed in Section 3.2.1.3, these techniques were not implemented in KE-PT, since frequency counts were sufficient. This also meant that the similarity judgments used in [39] were unnecessary. The key element implemented in KE-PT from [39] was the idea of having participants choose which moral principle they would employ to resolve a dilemma in TE form and using their responses to glean information about their moral beliefs. It is the efficacy of this procedure that is called into doubt by the performance of KE-PT.

[5] This is because the modifications that were made to the MFQ questions used in HCI-PT, were not accounted for in the studies validating the scale [29].

[6] This is because the TEs used in it are not drawn from a psychometrically validated instrument.

6.9 Implications for multi-theory AMAs in general

While the present study focused on the design of a tool to accurately elicit a user's preferred moral theory, it does not attempt a comparison between these multi-theory approach and single theory approach in terms of user experience. Thus, even if the findings here and in [11] are correct in suggesting the lack of a single moral theory in lay-cognition, the findings here do not invalidate the Genet project. This is because one way of understanding what Genet aims to do is maximize alignment between the user and the system. Here alignment can be operationalized as how often, if put in the exact same situation, the AMA acts as the user would (or at least in a way that is compatible the user's moral theory) [1]. The present project and [11] both observed some statistically significant correlations, which suggests that what some moral psychologistscall proto-moral theories [11] do seem to play a role in moral judgment. Even if these proto- theories are only one factor that contributes to moral judgment and action. This suggests that successfully detecting these theories and having an AMA act accordingly would increase the alignment by some degree over that offered by a one-theory-only approach. This is because these would only reduce user-machine misalignment for those who have that one proto-theory, while Genet would reduce this misalignment for all theories it covers. This in turn would lead to a better user experience for multi-theory system users. While this is a plausible hypothesis, there have yet to be empirical investigations addressing it.

6.10 Summary

The main explanatory possibilities for the observed performance of the prototypes explored in this chapter have been 1) the small sample size 2) measurement error and/or 3) the impossibility of the task given the nature of human moral psychology, in that non-philosophers do not reliably act/make judgements according to a single moral theory (or at least not the moral theories the prototypes attempt to detect).

The observation of unpredicted correlations and few statistically correlations in the inter-validation correlation analysis suggests that the performance observed in the initial correlational analyses was not an artefact generated by the measurement errors.

These findings, along with the mostly unacceptable CAs, suggest the results are due to contradictory answers given by participants, in addition to a sample size, which (even when using all the participants together in the intra-validation analyses) is insufficient to detect small correlations which are also significant.

It thus seems the small but significant correlations observed in [11] more accurately reflect the effect size in the actual population (as evidenced by the reduced correlation sizes in inter-validation analysis).

If this is true, then this would be a significant obstacle to any elicitation mechanism which aims to elicit a single preferred moral theory from users. This is because the evidence collected here, as well as the evidence presented by moral psychologists [11], suggests that the most likely explanation is that, in the population of non-philosophers, people do not have one preferred moral theory.

In terms of the implications of this for the comparison of moral psychological, HCI, and expert systems approaches, the failure to reject H1 implies none of them do better than the others. They thus are roughly equally (in)effective at accurately eliciting users' preferred moral theories.

However, this still leaves open the possibility that such a multi-theory, user-configurable system performs better than single theory systems (at least by a small degree). More work is thus needed to ascertain whether users do have a single moral theory, how best to elicit this theory, and how multi-theory systems compare to single-theory systems.

6.11 Limitations

Firstly, there were small sample sizes for each of the prototypes. This is a limitation, as it makes the observed statistically significant correlation coefficients more vulnerable to Type I error than in studies with larger sample sizes.

There were also a few measurement errors, which were extensively acknowledged in Chapter 5. As has been discussed this chapter though, given the inter-validation correlation and intra-validation CAs show a majority of statistically insignificant and unreliable results as well, it is not clear that these errors significantly affected the other results observed.

In addition to the above, there was also a limited number of moral theories investigated. This leaves open the possibility that the results observed are due to the participants not subscribing to any of the moral theories investigated, but to some other moral theory, rather than no theory at all, nor to a mixture of different theories.

The study also did not control for cultural and other demographic variables. This is an important issue, since the scales used in the design of the prototypes (particularly CogSci-PT) were validated with North American and European populations [11][14][19][30]. Thus, the results observed in the present study could be reflective of the invalidity of these scales for South African populations.

Finally, since the sampling in this project was non-random, this could result in its findings reflecting the peculiarities of an unrepresentative sample, rather than of actual phenomena present in the population of interest.

7. CONCLUSION

This chapter starts with a summary which restates the questions and the strategy employed for answering them, before providing answers to them. It then summarizes the implications of these answers. The contributions made by the project are then addressed. Finally, suggestions for future work are discussed.

7.1 General conclusion

This study set out to answer the questions:

> 1) Which of the three approaches to user preference elicitation (that of HCI, cognitive science, knowledge engineering/elicitation) as instantiated in high fidelity prototypes delivers the highest validity and reliability in its ability to elicit users' moral beliefs at Genet's base-theory level?
>
> 2) Which of the three approaches to user preference elicitation (that of HCI, cognitive science, knowledge engineering/elicitation) as instantiated in high fidelity prototypes is the most highly rated in terms of usability?

Three high fidelity prototypes were thus designed and implemented. Each representing one of the three disciplines, to observe how the approaches performed.

At least for the approaches and prototypes as instantiated in this study, the answer to 1) is that none of the prototypes perform better than the others according to the comparison of the magnitude of the statistically significant correlations observed between the prototype-assigned scores and the validation measure scores, as discussed Chapters 5 and 6. However, HCI-PT shows the most potential, as it was able to detect one statistically significant correlation (the same as CogSci-PT), and produced no correlations with unpredicted directionality. The answer to 2) is also that none of the prototypes perform statistically significantly differently than the others in terms of usability as measured on the SUS.

These findings, especially when consider in conjunction with [11], suggest that in the population of non-philosophers people do not have one preferred moral theory. If this is true, then it would be a significant obstacle to any elicitation mechanism which aims to elicit a single preferred moral theory from users. However, this still leaves open the possibility that such a system performs better. More work is thus needed to ascertain whether users do have a single moral theory, how best to elicit this theory and how these user configurable AMA compare to single-theory ones.

7.2 Contributions

As the first study to compare different moral theory elicitation approaches, the present study contributes to evidence for (or at least fails to falsify) problems with the project of making the design of AMAs dependent on elicitation of a user's one preferred moral theory. A positive claim that the data collected here does support is that, at least for some potential users, even computational elicitation tools that use validated measures of moral theory preferences do not allow one to predict the moral judgements they will make with high confidence.

In terms of the broader significance of these findings, the current project can be seen as an exemplar of a deeper interaction between moral psychology, moral philosophy and elicitation techniques drawn from various subdisciplines in computer science, namely HCI (particularly

value sensitive design) and knowledge engineering (particularly knowledge elicitation and knowledge-based AMA design). While, as noted in Section 2.2, work such as [10] does much to make these connections salient, this work does not show the relevance of existing psychometric instruments developed by moral psychologists (sometimes in tandem with moral philosophers) to answering questions that AMA research raises. The current project does this, by showing how research in the computer scientific subdisciplines can be translated into computational tools with the help of moral psychological scales, which can be used to answer the questions: "Can the moral theory that a user subscribes to be identified in an automated way? And, if so, how?".

7.3 Suggestions for future research

One avenue for future research to explore in light of the present project is 1) attempting to replicate the findings here with a larger sample, and 2) comparing these multi-theory prototypes to single theory prototypes to observe the effect of the degree of alignment on the user experience of the system. Conducting studies to address 1) will shed light on whether the small sample size does account for the findings of this project in the way suggested in Chapter 6. While 2) will reveal whether the tools developed here are better predictors of user moral judgment than single theory tools. It would also show whether, if there is an increase in alignment, this leads to an increase in the quality of the user experience.

Another avenue for future research is the investigation of alternate classification rules for moral theory classification other than those used in the prototypes tested here. For instance, while all prototypes in this study used straight forward comparisons for classification (e.g. if $x > y$, where "x" is a moral theory score based on user responses and "y" is some threshold value), rules that use a qualified comparison, such as one where an inequality like ">" could be taken to hold only if the compared values differ for a given gap (e.g. x is higher than y only if $x > y + z$, where $z > 0$ is the gap) could also be investigated. These rules would have the added advantage of allowing the trustworthiness of the classification to be quantified via the gap value.

Future work should also investigate prototypes with a larger number of moral theories than was investigated here (i.e., four). This research should also investigate the effect of sample demographics on moral theory elicitation. Finally, these future studies suggested above should also make use of random sampling procedures.

8. BIBLIOGRAPHY

[1] G. Rautenbaugh and M. Keet, “Toward equipping artificial moral agents with multiple ethical theories,” in *Proceedings of RobOntics: international workshop on ontologies for autonomous robotics*, D. Beßler, S. Borgo, M. Diab, A. Gangemi, A. Olivares, M. Pomarlan, and R. Porzel, Eds., Bozen-Bolzano: CEUR-WS, Sep. 2020.

[2] M. Anderson and S. L. Anderson, “Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm,” *Industrial Robot: An International Journal*, 2015.

[3] M. Anderson and S. L. Anderson, “GenEth: A general ethical dilemma analyzer,” *Paladyn*, vol. 9, no. 1, pp. 337–357, 2018.

[4] D. Vanderelst and A. Winfield, “An architecture for ethical robots inspired by the simulation theory of cognition,” *Cogn Syst Res*, vol. 48, pp. 56–66, 2018.

[5] B. Liao, M. Slavkovik, and L. van der Torre, “Building jiminy cricket: An architecture for moral agreements among stakeholders,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 147–153.

[6] J. E. Edlund and A. L. Nichols, Eds., *Advanced Research Methods for the Social and Behavioral Sciences*. Cambridge: Cambridge University Press, 2019.

[7] K. DeLapp, “Metaethics | Internet Encyclopedia of Philosophy,” *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/metaethi/

[8] A. Ecoffet and J. Lehman, “Reinforcement Learning Under Moral Uncertainty,” *proceedings.mlr.press*, Jul. 01, 2021. https://proceedings.mlr.press/v139/ecoffet21a.html (accessed May 30, 2024).

[9] B. Malle and M. Scheutz, “Moral competence in social robots,” in *In Machine ethics and robot ethics*, W. Wallach and P. Asaro, Eds., Routledge, 2020, pp. 225–230.

[10] E. Awad *et al.*, “The moral machine experiment,” *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[11] G. Kahane *et al.*, “Beyond sacrificial harm: A two-dimensional model of utilitarian psychology.,” *Psychol Rev*, vol. 125, no. 2, p. 131, 2018.

[12] R. N. Barger, *Computer ethics: A case-based approach*. Cambridge University Press, 2008.

[13] G. Gutting, *Thinking the impossible: French philosophy since 1960*. Oxford University Press, 2011.

[14] A. Simpson, J. Piazza, and K. Rios, “Belief in divine moral authority: Validation of a shortened scale with implications for social attitudes and moral cognition,” *Pers Individ Dif*, vol. 94, pp. 256–265, 2016.

[15] C.-K. Pai, R. Lee, D. Hinds, W. Xia, and B. Seaton, “How Do People Resolve Dilemmas? Eliciting Subjective Decision Factors,” in *2011 44th Hawaii International Conference on System Sciences*, IEEE, 2011, pp. 1–10.

[16] J. Prinz, “The emotional basis of moral judgments,” *Philosophical Explorations*, vol. 9, no. 1, pp. 29–43, 2006, doi: 10.1080/13869790500492466.

[17] K. Tobia, W. Buckwalter, and S. Stich, “Moral intuitions: Are philosophers experts?,” *Philos Psychol*, vol. 26, no. 5, pp. 629–638, 2013.

[18] J. Piazza and J. Landy, “‘Lean not on your own understanding’: belief that morality is founded on divine authority and non-utilitarian moral thinking,” *Judgm Decis Mak*, vol. 8, no. 6, pp. 639–661, 2013.

[19] R. H. Weigel, D. J. Hessing, and H. Elffers, “Egoism: Concept, measurement and implications

for deviance," *Psychology, Crime and Law*, vol. 5, no. 4, pp. 349–378, 1999.

[20] J. Robinson, "The Consequentialist Scale: Elucidating the Role of Deontological and Utilitarian Beliefs in Moral Judgments." Available: https://tspace.library.utoronto.ca/bitstream/1807/33868/3/Robinson_Jeffrey_S_201211_MA_thesis.pdf

[21] P. Conway, J. Goldstein-Greenwood, D. Polacek, and J. D. Greene, "Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers," *Cognition*, vol. 179, pp. 241–265, 2018.

[22] T. Molose, P. Thomas, and G. Goldman, "A qualitative approach to developing measurement scales for the concept of Ubuntu," *Acta Commercii*, vol. 19, Aug. 2019, doi: 10.4102/ac.v19i1.692.

[23] C. H. McCurrie, D. L. Crone, F. Bigelow, and S. M. Laham, "Moral and Affective Film Set (MAAFS): A normed moral video database," *PLoS One*, vol. 13, no. 11, p. e0206604, 2018.

[24] B. Friedman, D. G. Hendry, and A. Borning, "A survey of value sensitive design methods," *Foundations and Trends in Human-Computer Interaction*, vol. 11, no. 2, pp. 63–125, 2017.

[25] P. Pu, B. Faltings, and M. Torrens, "User-involved preference elicitation," 2003.

[26] E. Niforatos, A. Palma, R. Gluszny, A. Vourvopoulos, and F. Liarokapis, "Would you do it?: enacting moral dilemmas in virtual reality for understanding ethical decision-making," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–12.

[27] S. A. Munson, D. Avrahami, S. Consolvo, J. Fogarty, B. Friedman, and I. Smith, "Attitudes toward online availability of US public records," in *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, 2011, pp. 2–9.

[28] L. P. Nathan, P. V Klasnja, and B. Friedman, "Value scenarios: a technique for envisioning systemic effects of new technologies," in *CHI'07 extended abstracts on Human factors in computing systems*, 2007, pp. 2585–2590.

[29] J. Graham, B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto, "Mapping the moral domain.," *J Pers Soc Psychol*, vol. 101, no. 2, p. 366, 2011.

[30] S. Clifford, V. Iyengar, R. Cabeza, and W. Sinnott-Armstrong, "Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory," *Behav Res Methods*, vol. 47, no. 4, pp. 1178–1198, 2015.

[31] D. L. Crone and S. M. Laham, "Multiple moral foundations predict responses to sacrificial dilemmas," *Pers Individ Dif*, vol. 85, pp. 60–65, 2015.

[32] J. J. Thomson, "The Trolley Problem," *The Yale Law Journal*, vol. 94, no. 6, pp. 1395–1415, May 1985, doi: https://doi.org/10.2307/796133.

[33] E. Machery, *Philosophy within its proper bounds*. Oxford University Press, 2017.

[34] N. R. Shadbolt, P. R. Smart, J. Wilson, and S. Sharples, "Knowledge elicitation," *Evaluation of human work*, pp. 163–200, 2015.

[35] M. Keet, *An introduction to ontology engineering*, vol. 1. Maria Keet Cape Town, 2018.

[36] E. Hudlicka, "Requirements elicitation with indirect knowledge elicitation techniques: comparison of three methods," in *Proceedings of the Second International Conference on Requirements Engineering*, IEEE, 1996, pp. 4–11.

[37] G. Kelly, "Personal construct psychology," *Nueva York: Norton*, 1955.

[38] J. M. Bradshaw, K. M. Ford, J. R. Adams-Webber, and J. H. Boose, "Beyond the repertory grid: new approaches to constructivist knowledge acquisition tool development," *International*

Journal of Intelligent Systems, vol. 8, no. 2, pp. 287–333, 1993.

[39] S. Verheyen and M. Peterson, "Can we use conceptual spaces to model moral principles?," *Rev Philos Psychol*, vol. 12, no. 2, pp. 373–395, 2021.

[40] G. Nicolas, X. Bai, and S. T. Fiske, "Comprehensive stereotype content dictionaries using a semi-automated method," *Eur J Soc Psychol*, vol. 51, no. 1, pp. 178–196, 2021.

[41] S. Ullah, M. Iqbal, and A. M. Khan, "A survey on issues in non-functional requirements elicitation," in *International Conference on Computer Networks and Information Technology*, IEEE, 2011, pp. 333–340.

[42] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, and H. M. Abbasi, "A survey of ontology learning techniques and applications," *Database*, vol. 2018, 2018.

[43] W. Chergui, S. Zidat, and F. Marir, "An approach to the acquisition of tacit knowledge based on an ontological model," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 7, pp. 818–828, 2020.

[44] M. Davis and R. S. Beidas, "Refining contextual inquiry to maximize generalizability and accelerate the implementation process," *Implement Res Pract*, vol. 2, p. 2633489521994941, 2021.

[45] E. Costetchi, "Towards a Discourse Model for Knowledge Elicitation," in *Proceedings of the Student Research Workshop associated with RANLP 2013*, 2013, pp. 38–44.

[46] E. Costetchi, "Towards Automated Ontology Elicitation Dialogues," *University of Luxembourg, Luxembourg, Master Thesis*, 2010.

[47] K. Data, "How to Calculate b0, b1, and b2 Coefficient Manually in Multiple Linear Regression," *KANDA DATA*, Mar. 29, 2022. https://kandadata.com/how-to-calculate-bo-b1-and-b2-coefficient-manually-in-multiple-linear-regression/

[48] P. Singer, "Famine, affluence, and morality," in *Applied Ethics*, Routledge, 2017, pp. 132–142.

[49] J. Brooke, "SUS: A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, Nov. 1995.

[50] D. C. Crocker, "Some Interpretations of the Multiple Correlation Coefficient," *Am Stat*, vol. 26, no. 2, pp. 31–33, 1972, doi: 10.2307/2683460.

[51] N. C. Silver and W. P. Dunlap, "Averaging correlation coefficients: Should Fisher's z transformation be used?," *Journal of applied psychology*, vol. 72, no. 1, p. 146, 1987.

[52] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[53] A. A. Agbo, "Cronbach's alpha: Review of limitations and associated recommendations," *Journal of Psychology in Africa*, vol. 20, no. 2, pp. 233–239, 2010.

[54] D. Bourget and D. Chalmers, "Philosophers on philosophy: The 2020philpapers survey," *Philosophers' Imprint*, vol. 23, no. 1, 2023.

APPENDICES

In the following sections are the elicitation stimuli used in the prototypes (in Appendix A) and material used in the validation surveys (in Appendix B). Note that in both Appendices, the sources of the material are cited according to their numbering in the bibliography in Chapter 8.

A.1 APPENDIX A

A.1.1 CogSci-PT scales

Below are the scales used in CogSci-PT.

A.1.1.1 ES questions

The questions below were drawn (without modification) from [19]. All of these questions required a response on a Likert scale of 1 (no agreement) -5 (maximal agreement).

1. All in all it is better to be humble and honest than important and dishonest.
2. A person should obey the law no matter how much it interferes with their own ambition.
3. Never tell anyone the real reason you did something unless it's useful to do so.
4. It's hardly fair to bring children into the world today with the way things look for the future.
5. It is always alright to 'bend' the law as long as you don't violate it.
6. It is hard to get ahead without cutting corners here and there.
7. Nowadays a person has to live just for today and let tomorrow take care of itself.
8. A person should obey only those laws that seem reasonable.
9. The best way to handle people is to tell them what they want to hear.
10. A person is justified in given false testimony to protect a friend on trial.
11. Most people don't care what happens to the next fellow.
12. Generally speaking, people won't work hard unless forced to do so.
13. Laws are so often made for the benefit of small, selfish groups that a person cannot respect the law.
14. Next to health, money is the most important thing in life.
15. The biggest difference between criminals and other people is that criminals are stupid enough to get caught.
16. You sometimes can't help but wonder if anything is worthwhile.
17. General rules about lying are useless, whether a lie is moral or immoral depends on the situation.
18. To make money, there are no right or wrong ways anymore, just easy and hard ways.
19. These days a person doesn't really know whom they can count on.
20. It is okay to break the law if you don't get caught.
21. This is a control question to check whether you are paying attention. Please proceed by selecting 1 on the scale below.

A.1.1.2 MFDA Questions

The questions below were drawn (without modification) from [14]. These questions required a response on a Likert scale of 1 (no agreement) -9 (maximal agreement).

1. Everything we need to know about living a moral life God has revealed to us.
2. What is morally good and right is what God says is good and right.
3. If you want to know how to live a moral life you should look to God.
4. Acts that are immoral are immoral because God forbids them.
5. It is possible to live a righteous life without knowledge of God's laws.
6. This is a control question to check whether you are paying attention. Please proceed by selecting 3 on the scale below.

A.1.1.3 OUS Questions

The questions below were drawn (without modification) from [11]. These questions required a response on a Likert scale of 1 (no agreement) -7 (maximal agreement)

1. If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.
2. It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.
3. From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.
4. If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.
5. From a moral perspective, people should care about the well-being of all human beings on the planet equally, they should not favor the well-being of people who are especially close to them either physically or emotionally.
6. It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.
7. It is just as wrong to fail to help someone as it is to actively harm them yourself.
8. It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal.
9. This is a control question to check whether you are paying attention. Please proceed by selecting 1 on the scale below.

A.2 HCI-PT Value scenarios

Below are the value scenarios used in HCI-PT, segmented by value the questions measure. These were all based on the MFQ-based vignettes in [30], and modified according to the specifications in [24] and [28]. All the value scenarios below required a response on a Likert scale of 1 (no agreement) -5 (maximal agreement)

A.2.1 Care

1. You live in a future where a new video streaming app becomes pervasive. The app allows creators to upload video ideas, with the users bidding on the ideas they would most like to see. The idea with the highest bid is then produced. On this app you come across 30-minute-long video of a dog outside alone in the rain as punishment for scratch in the garbage.
2. You live in a future where a new pet-centered social media app becomes pervasive. The site allows users to post pictures and videos of their pets. On this site you see a video of a man lashing his pony with a whip for breaking loose from its pen.
3. You live in a future where a new video streaming app becomes pervasive. Because of the app's various educational features, it is especially popular with students and teachers. On this site you see a video of a teacher hitting a student's hand with a ruler for falling asleep in class.
4. You live in a future where a new dating app becomes pervasive. This dating app allows one to set up blind dates with users in their area without seeing their picture. One day, at a restaurant, you sit behind a man who you see using the app on his phone. As a woman enters the restaurant you see the man look at her face and rapidly cancel the date
5. This is a control question to check whether you are paying attention. Please proceed by selecting 3 on the scale below.
6. You live in a future where a new parenting-centered social media app becomes pervasive. On this site, you see a woman spanking her child with a spatula for getting bad grades in school.
7. You live in a future where a new bus seat booking app that allows one to see pictures of those who have also booked becomes pervasive. One day you take a bus without using the app and notice that no one has booked either seat next to an obese woman.

A.2.2 Fairness

1. You live in a future where an automated study assistant which answers students' school-work related questions becomes pervasive. Students are not allowed to use the assistant during tests and for the most part do not do so. One day You see a student using the assistant during a test.
2. You live in a future where a soccer videogame becomes pervasive. This game allows players more control over their avatar's behaviour, such as allowing them to trigger a reaction as if their avatar had gotten fouled even if they hadn't. One day you see a player making his avatar pretend to be seriously fouled by an opposing player.
3. You live in the future where a whistle-blower app has become pervasive on university campuses. The app allows students and other staff to anonymously report misconduct directly to university administration. Unfortunately, one professor breaks the encryption and is able to find the name of a student who reported him. One day you see this professor giving a bad grade to the student who reported him just because he dislikes him.
4. This is a control question to check whether you are paying attention. Please proceed by selecting 2 on the scale below.
5. You live in a future where a soccer videogame becomes pervasive. The game allows players to take control of the referee. One day in a competitive match of this game,

played for money, you see a referee intentionally making bad calls that help his favoured team win.
6. You live in a future where a line booking app becomes pervasive. The app allows one to book a spot in most queues before one arrives in the queue. One day you see a boy skipping to the front of the line by using his friend's employee code in the app

A2.3 Authority

1. You live in a future where a soccer videogame becomes pervasive. The videogame allows players to inhabit the coach as well in the players on the field. In a competitive match of this game, played for money, you see a player publicly yelling at his soccer coach during the game.
2. You live in a future where a new education-centred video streaming site becomes pervasive In one of the videos of this platform you see a teaching assistant talking back to the teacher in front of the classroom.
3. You live in a future where religious-service video streaming site becomes pervasive. Now, only a minority of people attend religious services in person. In one of the videos on this platform, you see a group of women having a long and loud conversation during a church sermon.
4. You live in a future where a soccer videogame becomes pervasive. The videogame allows players to inhabit the coach as well in the players on the field. In a competitive match of this game, played for money, you see a star player ignoring her coach\'s order to come to the bench during a game.
5. This is a control question to check whether you are paying attention. Please proceed by selecting 4 on the scale below.
6. You live in a future where an automated study assistant which answers students' school-work related questions becomes pervasive. One day you see a girl using the assistant to learn the answers before her teacher can properly explain it then repeatedly interrupting her teacher as he tries to explain the concept.

A2.4 Loyalty

1. You live in a future where a soccer videogame becomes pervasive. The videogame allows players to inhabit the coach as well in the players on the field. In a competitive match of this game, played for money, you see a coach celebrating with the opposing team's players who just won the game.
2. You live in a future where school level competitive contest video streaming site becomes pervasive. On this site you see a teacher publicly saying she hopes another school.
3. You live in a future where an app that allows video streaming of beauty pageants becomes pervasive. This app also allows viewers to vote for a winner. One day you see a man using the app and secretly voting against his wife in a beauty pageant.
4. You live in a future where a political news video streaming site becomes pervasive. On this site, you see the SA Ambassador joking in Great Britain about the stupidity of South Africans.

5. You live in a future where a political news video streaming site becomes pervasive. On this site, you see a former South African minister publicly giving up his South African citizenship.

A2.5 Sanctity

1. You live in a future where a dare video streaming site becomes pervasive. The site allows users to post dares ideas and other users to upvote those ideas. On this site you see a man in a bar using his phone to watch people having sex with animals.
2. You live in a future where a dare video streaming site becomes pervasive. The site allows users to post dares ideas and other users to upvote those ideas. On this site you see an employee at a morgue eating his pepperoni pizza off of a dead body.
3. You live in a future where a dare video streaming site becomes pervasive. The site allows users to post dares ideas and other users to upvote those ideas. On this site you see a man searching through the trash to find women's discarded underwear.
4. You live in a future where humanoid robots become pervasive. One day you see a single man ordering such a robot who looks like his secretary to have sex with.
5. You live in a future where a dare video streaming site becomes pervasive. The site allows users to post dare ideas and it allows other users to upvote those ideas. Users then complete the dares with the highest number of upvotes. Once a dare is completed, a video of the completion is uploaded to the site. In a video on this site, you see a man having sex with a frozen chicken before cooking it for dinner.

A.3 KE-PT TES AND PRINCIPLES

Below are the thought experiments displayed to the user by KE-PT, and the principles they could choose in response. Where the content was taken from a pre-existing source, this source is cited next to the content itself.

A3. 1 TEs

1. You are standing next to a fork in a train track and a switch to divert a train that is about to kill five workers unless you throw the switch and divert the trolley down a side track where it will kill one worker. [32]
2. Lisa gets a call from the bank telling her that hackers have hacked into her bank account. They are drawing $200 every minute. Lisa hurries to the bank to sort this out. On her way there, she realizes that children are drowning in a pond. No one else is there to help. Lisa took a while to calculate how many children she could save before she's bankrupt. She can only save 30, leaving the rest of the children to drown. [48]
3. One day you accidentally stumble upon a death cult. The cultists tell you that if you don't renounce your God (if you have one) and accept their god, Mictlantecuhtli, they will kill a child that they've kidnapped.
4. One day your friend tells you that he is sleeping with a woman who is not his wife. You could stop this by telling his wife, but if you do this it will end his marriage, leaving their young daughter Alice to grow up in a broken home.
5. One day a man knocks on the door. The man says that he has taken your neighbour, whom you are close with, hostage. He goes on to say that if you admit to being friends

with your neighbour, he will release your neighbour and take you hostage instead. If you don't say your neighbour is your friend, he will kill them.

6. Imagine a man named Vuyani. Vuyani has to choose between going to his estranged sister's birthday or studying for a test. He has had a poor relationship with his sister for the last 20 years, and this is the first time he's been invited to her birthday since they were children (despite them still belonging to the same community). If he doesn't go to the sister's birthday, he might not get another chance to better their relationship. But if he fails the test, he will probably not be granted the PhD his been pursuing.
7. This is a control question to check whether you are paying attention. Please proceed by selecting: "An action is right if out of its alternatives it is the one that increases overall happiness at the smallest emotional cost. on the scale below.

A3.2 Principles

1. An action is right if out of its alternatives it is the one that increases overall happiness at the smallest emotional cost. (Utilitarianism)
2. An action is right if its ends are in the agent's self-interest (Egoism)
3. An action is right if God wills it so (DCT)
4. An action is right if it is universally willable and it respects the rational autonomy of others. (Deontology)

APPENDIX B

This appendix contains the questions asked as part of the validation survey participants took after interacting with one of the prototypes. Besides the SUS scale, which is drawn from [49], where content was taken from a pre-existing source, this source is cited next to the content itself.

B.1. PRELIMINARY QUESTIONS

1. What participant number were you given in the previous section of the study?
2. What did the prototype you interacted with classify you as?
3. Do you have at least a master's degree in philosophy?
4. Is your area of specialization moral philosophy?
5. Which prototype did you interact with?
6. What would you say your ethical outlook is?

B.2. VALIDATION SCENARIOS

1. Imagine you are approached by a known killer and he asks you to give the address of one of your colleagues, and you have good reason to believe that once he acquirers the address he will kill the colleague in question. You definitely know the correct address. On a scale of 1 to 5 (with 1 being not wrong at all, and 5 being very wrong), how wrong would it be to lie and give the wrong address? (Deontology)
2. You are standing on a footbridge that is going over the train tracks and something similar to the scenario above is happening – a trolley is heading towards five people. You can push someone off the footbridge where you're standing to hit the carriage and the trolley will stop, saving the five people's lives but killing the person you pushed. On a scale of 1 to 5 (with 1 being not wrong at all, and 5 being very wrong), how wrong would pushing the person be? (Deontology/utilitarianism) [32]
3. Imagine you've promised your friend that you will meet them for an important event that cannot be rescheduled. On the day you are supposed to keep this promise, you win free tickets to see your favourite musician live at the same time as your friend's event. On a scale of 1 to 5 (with 1 being not wrong at all, and 5 being very wrong), how wrong would it be for you to attend the concert and not your friend's event? (Egoism)
4. Imagine that a friend invites you to attend a religious service with her. In this service, everyone in attendance is forced to worship an idol or have a finger cut off. On a scale of 1 to 5 (with 1 being not wrong at all, and 5 being very wrong), how wrong would it be for you to worship the idol in order to avoid losing a finger? (DCT)
5. Imagine you come into contact with a genie who says she can either give you unlimited wealth, or she can end world poverty. On a scale of 1 to 5 (with 1 being not wrong at all, and 5 being very wrong), how wrong would it be for you to ask for unlimited wealth instead of an end to world poverty? (Utilitarianism)
6. Imagine you have a job interview for your dream job. The interview goes well, but the last thing you required to do is take the lord's name in vain. If you don't do this, you will not get the job and you have good reason to believe that you will be blacklisted by

the entire industry. On a scale of 1 to 5 (with 1 being not wrong at all, and 5 being very wrong), how wrong would it be for you take the lord's name in vain in order to get the job? (Egoism)

7. Imagine you are running a marathon, where the winner gets R1 million. You are in the lead. When you are a few hundred metres from the finish line, you see a woman being mugged in the ally. You are certain you could save the woman, but you are also certain that it would mean you would lose your first place spot. On a scale of 1 to 5 (with 1 being not wrong at all, and 5 being very wrong), how wrong would it be for you to ignore the mugging to retain your winning placement? (Egoism)

B.3. VALIDATION EXPLICIT STATEMENTS

1. The only thing that determines whether an act is morally right is whether, out of the available options, it is the act that would lead to the most happiness and the least suffering in the world, taking into account the welfare of all sentient beings, whether human or animal. An act that doesn't maximize welfare in this way is morally wrong. On this moral view, no one counts for more than anyone else: our own interests and needs, and the interests and needs of our family and friends, never count for more than the interests and needs of any other person, however distant from us. Finally, on this view the only thing that matters is how our actions affect the amount of happiness in the world. It is always morally right to break a rule or principle if doing so would lead to the better outcome. On a scale of 1 to 5 (with 1 being no agreement at all, and 5 being totally agree), how much do you agree with the statement above? (Utilitarianism) [11]
2. God decrees what is right and wrong. Actions that God commands us to do are morally required; actions that God forbids us to do are morally wrong; and all other actions are morally neutral. So, something is right or wrong is perfectly objective: It is right if God commands it and wrong if God forbids it. On a scale of 1 to 5 (with 1 being no agreement at all, and 5 being totally agree), how much do you agree with the statement above? (DCT) [14]
3. Each person ought to pursue his or her own self-interest exclusively. So, what makes an action right is whether it is in one's self-interest, and when an action is not in one's best interest, it is wrong. Thus, if we are to help others, the benefit to others is not what makes the act right. Rather, the act is right because it benefits you. This does not imply that in pursuing your interests, you should always do what you want to, or what offers you the most short-term pleasure. Someone may want to smoke cigarettes, or bet all his money at the racetrack, or set up a meth lab in his basement. These actions are not right, despite their possible short-term benefits. A person ought to do what really is in his or her own best interests, over the long run. On a scale of 1 to 5 (with 1 being no agreement at all, and 5 being totally agree), how much do agree with the statement above? (Egoism)[19]
4. Some choices cannot be justified by their effects—that no matter how morally good their consequences, some choices are morally forbidden. A person cannot make certain wrongful choices even if by doing so the number of other people who will make those exact kinds of wrongful choices will be reduced. What makes a choice right is that it is consistent with a moral rule. Such rules are to be simply obeyed by each person. If an act is not in accord with the moral rules, it may not be undertaken, no matter the

good outcomes that it might produce. This is the case, even though these rules do not derive from god's commands. On a scale of 1 to 5 (with 1 being no agreement at all, and 5 being totally agree), how much to you agree with the statement above? (Deontology)[11]

B.4. SUS SCALE

1. I think I would like to use the prototype frequently.
2. I found the prototype unnecessarily complex.
3. I thought the prototype was easy to use.
4. I think that I would need the support of a technical person to be able to use the prototype.
5. I found the various functions in the prototype were well integrated.
6. I found the various functions in the prototype were well integrated.
7. I thought there was too much inconsistency in the prototype.
8. I would imagine that most people would learn to use the prototype very quickly.
9. I found the prototype very cumbersome to use.
10. I felt very confident using the prototype. I needed to learn a lot of things before I could get going with this prototype.
11. How likely are you to recommend this website/prototypes to others? 1 represents "not likely at all" and 10 represents "extremely likely".